Essays on Ethics in Economics

Dissertation
submitted to the Faculty of Economics,
Business Administration and Information Technology
of the University of Zurich

to obtain the degree of
Doktor der Wirtschaftswissenschaften, Dr. oec.
(corresponds to Doctor of Philosophy, PhD)

presented by

Florian Engl
from Germany

approved in September 2015 at the request of

Prof. Dr. Roberto A. Weber
Prof. Dr. María Sáez Martí

The Faculty of Economics, Business Administration and Information Technology of the University of Zurich hereby authorizes the printing of this dissertation, without indicating an opinion of the views expressed in the work.

Zurich, 16.09.2015
Chairman of the Doctoral Board: Prof. Dr. Josef Zweimüller

Acknowledgements

I want to thank my supervisor, Prof. Roberto Weber, for excellent guidance and support throughout my PhD studies, and my co-supervisor, Prof. María Sáez Martí, for numerous helpful discussions and encouragements.

My co-authors, Prof. Björn Bartling, Prof. Arno Riedl and Prof. Roberto Weber, were, and continue to be, role models for me and I want to express my gratitude for giving me the opportunity to learn so much from them.

I am deeply grateful to all my colleagues and friends for making my time in Zurich joyful and inspiring, and, especially, to my girlfriend, Lea Cassar, for her great patience and help during the difficult times of the PhD studies. Finally, I wish to thank my parents. Without their support throughout my studies this project would not have been possible.

Contents

Chapter 1

Introduction

The main focus of my thesis lies on gaining a better understanding of the mechanisms that guide the attribution of blame and praise. Understanding why we judge some behavior as more blame- or praiseworthy than others is of great importance for ethics and economics. First, because it allows us to study how people make value judgments in economically important situations and second, because the attribution of blame and praise influences the behavior of those subject to it.

In the second chapter of my thesis, my coauthors, Prof. Björn Bartling and Prof. Roberto A. Weber, and I experimentally answer the question of whether people can avoid blame by remaining willfully ignorant about possible negative consequences of their actions for others. This research question was motivated by the observation that, in corporate and political contexts, individuals often present ignorance as an excuse for why they should not be held responsible for adverse outcomes that their actions caused. The study directly addresses this issue, by quantifying the extent to which engaging in willful ignorance allows a decision maker to deflect external blame for his actions and their consequences. To this end, we conducted a laboratory experiment in which some participants could choose to remain ignorant about the consequences of their actions for others and in which other participants had the opportunity to impose costly monetary punishment after observing behavior and the resulting outcomes. We interpret the assigned punishment as a measure of blame for an action and its consequences. The results show that, when taking an action that increases one's own welfare, but also results in harm to others, it is better to have avoided knowledge that harm would occur. Thus, willful ignorance can help avoid blame. At the same time, however, we find that willful ignorance itself is evaluated negatively, regardless of the consequences. That is, choosing to forgo information concerning the consequences of one's action and acting in a self-regarding way incites blame, even when the resulting consequence is beneficial for others. By remaining ignorant the decision

maker shows disregard for the possibility that others may be harmed and this appears sufficient for provoking blame and punishment.

In the third chapter of my thesis, I study how perceptions of responsibility guide the attribution of blame and praise. Through this channel, responsibility perceptions play a crucial role in many environments that are of traditional interest to economists. For example, in labor markets, workers and managers are fired or rewarded depending on their responsibility for the failure or success of a project and in political economy contexts, people make voting decisions based on the attribution of responsibility to politicians for the implementation of reforms and economic outcomes. I develop a notion of *causal responsibility* that measures the causal importance of a person's action for the implementation of an event, when the event's implementation depends on the interaction of potentially many parties, persons and/or chance. I incorporate the notion in a framework of *responsibility preferences* in which agents value monetary payoff, but also have a taste to punish (reward) other agents for the implementation of what they judge are bad (good) events, but only to the extent that these agents are causally responsible for the event. Furthermore, I study how those subject to responsibility-driven attribution of punishment and reward react to it, thereby analyzing the consequences of responsibility perceptions for equilibrium outcomes in a game-theoretic environment. I show how, depending on the specific environment, causal responsibility perceptions can induce distinctively different equilibrium outcomes in which, in some cases, causal responsibility for an event is maximally diffused between all, and, in other cases, maximally focused on some of the involved agents. Finally, I test the predictive power of the causal responsibility notion for the allocation of punishment in data from existing, incentivized experiments. I find that it can explain observed punishment patterns in many cases more successfully than existing theories and that it remains a highly significant predictor for punishment even after controlling for several other potential punishment motives.

While chapters two and three study the attribution of blame and praise and therefore focus on the reaction to certain behavior, in the fourth chapter my coauthors, Prof. Arno Riedl and Prof. Roberto A. Weber, and I study how ethical behavior, and specifically cooperative behavior, itself is influenced by the constraints and institutions that are in place in a society. Institutions are an important means for fostering prosocial behaviors. For example, sanctioning institutions have been shown to be effective for supporting high levels of cooperation in social dilemmas. Moreover, institutions may directly shape individuals' preferences and beliefs. In many contexts, however, institutions are limited in scope and can govern prosocial behavior only in some domains. In other domains, society must rely on voluntary prosocial behavior of individuals. We use a laboratory experiment to study how the presence and nature of an institution that enforces prosocial behavior in one domain affect the behavior in other domains, beyond the reach of the institution. In

addition, we study if and how the presence of an institution alters prosocial preferences and beliefs about others' behavior. Groups play two identical public good games, with one game potentially governed by an institution enforcing cooperation. We vary whether the institution is absent, imposed exogenously, or arises endogenously through voting by group members. We find that the presence of an institution in one game generally enhances cooperation in the other game. However, cooperation boosted by an exogenously imposed institution nevertheless decays over time, while the endogenously determined institution leads to stable spillover effects on voluntary cooperation levels. We also find that the presence of an institution strengthens beliefs about others' prosocial behavior and enhances prosocial preferences even towards strangers. When deciding about the implementation of such institutions in reality, these effects should to be taken into account by policy makers and can, potentially, alter the analysis in favor of (endogenously implemented) institutions.

Chapter 2

Does Willful Ignorance Deflect Punishment? – An Experimental Study[1]

"A man is responsible for his ignorance." – Milan Kundera, *Laughable Loves*

2.1 Introduction

Many important decisions involve tradeoffs between personal benefits and impacts on the welfare of others. In such situations, there is often the possibility of remaining uninformed about how one's actions affect others. Moreover, such "willful ignorance" may provide a justification for self-interested behavior. That is, while a decision maker is typically held responsible for knowingly committing an action that hurts others, the attribution of responsibility is less clear when he acts without knowledge of consequences. Such reasoning may even hold when the decision to remain ignorant is made privately, as ignorance allows one to act selfishly without direct confrontation with the consequences for others or the associated guilt (Dana et al., 2007). Thus, strategically manipulating one's information about the consequences of one's actions for others provides a path through which ignorance, even when deliberate, might provide insulation from responsibility or blame.

[1] This paper is published as Bartling, Björn, Florian Engl, and Roberto A. Weber. "Does willful ignorance deflect punishment?–An experimental study," European Economic Review, 70, 512-524, 2014. We would like to thank an associate editor, two anonymous referees, Martin Dufwenberg, Bertil Tungodden, and participants at the ABEE Symposium 2012 on Behavioural Economics in Markets and Organizations in Amsterdam, the 2012 Zurich Workshop in Economics, the Sixth Annual NYU-CESS Conference on Experimental Political Science 2013, the 2013 Spring School in Behavioral and Experimental Economics in San Diego, and the 2013 Asia-Pacific ESA Conference in Tokyo for valuable comments. We gratefully acknowledge financial support from the Foundation for Research in Science and the Humanities at the University of Zurich.

In corporate and political contexts, individuals often present ignorance as an excuse for why they should not be held responsible for adverse outcomes that they caused. For example, following corporate scandals and fraud, CEOs and board members often excuse their role by claiming they were not aware of what took place further down the hierarchy. Examples include former Enron CEO Kenneth Lay, who claimed ignorance about any accounting irregularities at the failed firm, and Rupert Murdoch, who was directly accused of showing "willful blindness" concerning the phone-hacking practices at News Corporation.[2] In the political sphere, public officials often argue that being unaware of acts committed by subordinates should exonerate them from blame.[3] For example, in response to revelations about the NSA's widespread wiretapping of allied leaders' phones, high-ranking U.S. government officials claimed lack of knowledge that these surveillance practices were taking place.[4]

Prior research in economics demonstrates that decision makers seize upon strategies to act self-interestedly at the expense of others, when presented with opportunities for avoiding blame or responsibility.[5] An important but largely open question, however, is to what extent such strategies are, in fact, effective in deflecting blame.[6]

Our study directly addresses this issue, by quantifying the extent to which engaging in willful ignorance allows a decision maker to deflect external blame for his actions and their consequences. To this end, we conduct a laboratory experiment in which some participants can choose to remain ignorant about the consequences of their actions for others and in which other participants have the opportunity to impose costly monetary punishments after observing

[2] See http://www.businessweek.com/stories/2006-02-05/commentary-ken-lays-audacious-ignorance and http://www.guardian.co.uk/media/2012/may/01/phone-hacking-report-wilful-blindness

[3] In fact, political science has long recognized the ability to avoid blame as an important determinant of a politician's success (Weaver, 1986) and ignorance as a potential strategy to do so (McGraw, 1991).

[4] See http://online.wsj.com/news/articles/SB10001424052702304470504579162110180138036

[5] Some research demonstrates that decision makers hide behind uncertainty - both their own and that of others - about what outcomes will result or how such outcomes were produced in order to keep more money in a distributional context (Dana et al., 2007; Andreoni and Bernheim, 2009; Ockenfels and Werner, 2012). In some cases, this can even mean that people are willing to accept less money in order to forgo the opportunity to share and have the other person know that sharing could have taken place (Broberg et al., 2007; Dana et al., 2006; Lazear et al., 2012). Hamman et al. (2010) show that delegating distributive decisions to others similarly provides a justification for self-interested behavior. More generally, a growing literature on behavioral ethics (Treviño et al., 2006; Bazerman and Gino, 2012) seeks to identify factors that influence ethical conduct, often highlighting how contextual features can lead otherwise "good" people to feel licensed to act unethically (Mazar et al., 2008; Dana et al., 2012).

[6] Experimental research in economics has only recently started to investigate the effectiveness of blame-avoidance strategies. For example, Bartling and Fischbacher (2012) show that delegating a decision that can lead to an unfair allocation is an effective way to shift blame from oneself toward the person to whom the decision is delegated.

behavior and the resulting outcomes. We interpret the assigned punishment as a measure of blame and responsibility attribution for an action and its consequences.

More precisely, in our experiment a dictator plays a binary dictator game under one of two possible states of the world. The state of the world is chosen by a random device and determines whether an action that is personally beneficial for the dictator benefits or harms the receiver. The dictator can decide whether or not to learn the true state, and faces no cost for acquiring this information. The realized state is irrelevant for the dictator's payoffs, meaning that ignorance creates no uncertainty about the dictator's payoffs, but enables the dictator to remain ignorant about the effects of his action on others. Thus, our design affords the dictator the opportunity to remain willfully ignorant regarding the social consequences of his actions.

Our focus is not on the effects of willful ignorance *per se*, however (cf. Dana et al., 2007), but instead on the extent to which remaining willfully ignorant allows the dictator to avoid blame and responsibility when a bad outcome results for the receiver. Therefore, in our experiment a third party observes the actions of the dictator and the outcome of the game and decides whether and to what extent to punish the dictator for his behavior.

Our results show that, when outcomes detrimental to the receiver result, ignorance is indeed effective in reducing punishment. That is, when taking an action that increases one's own welfare, but also results in harm to others, it is better to have avoided knowledge that harm would occur. Thus, willful ignorance can help avoid blame.

At the same time, however, we find that willful ignorance itself is evaluated negatively, regardless of the consequences. That is, choosing to forgo information concerning the receiver's payoffs and acting in a self-regarding way incites punishment, even when the resulting state of the world is one in which the dictator's self-interest is also beneficial for the receiver. By remaining ignorant the dictator shows disregard for the possibility that the receiver may obtain a low payoff and this appears sufficient for inducing punishment by third parties. Thus, the mere act of avoiding information about how one's decisions affect others provokes blame and punishment.

As a result of the above two counteracting effects of ignorance on punishment by third parties, in expectation, willful ignorance does not yield a higher payoff than knowingly acting selfishly. That is, while ignorance provides some blame avoidance when bad outcomes result, the fact that its use produces blame even when the outcomes are good makes it an ineffective strategy for obtaining higher payoffs in our experiment.

However, the punishment *pattern* revealed in our study has important implications for how willful ignorance might interact with punishment outside the laboratory. Attention to the possibility of blame and punishment is often salient only when bad outcomes arise – e.g., following a scandal or harmful misdeed. The fact that decision makers are penalized less when acting under willful ignorance therefore suggests that willful ignorance may be a good strategy in contexts where punishment is unlikely to be considered absent some noticeably bad consequence. Thus, corporate and political leaders who suspect wrongdoing in the institutions they manage may, indeed, benefit from a strategy involving willful ignorance.

Our results also have implications for economic theories of social preferences. We find significant differences in punishment for the same outcome, depending on whether the dictator revealed the state before making his choice. This cannot be explained by theories that incorporate social motives through preferences over final payoff distributions (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000), which predict the same punishment for an allocation, independently of the actions that led to the allocation. The qualitative comparative-static effect of willful ignorance on punishment is consistent with theories that incorporate intention-based reciprocity as a motive (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006; Sebald, 2010). However, these models fail to predict our additional finding that willfully ignorant dictators are still punished less when the beneficial outcome occurs than when the unfair outcome obtains. That is, outcomes matter even for willfully ignorant dictators.

Research on procedural fairness recognizes that people care not only about distributions of final outcomes, but also about the procedures employed to implement outcomes (Frey et al., 2004; Bolton et al., 2005; Trautmann, 2009; Krawczyk, 2011; Fudenberg and Levine, 2012). Our study contributes to this literature in that we show that punishment is not determined solely by consequences, but also by the process – in our case, the dictator's decision whether to acquire information – that leads to those consequences. Our research thus also relates to recent studies that find both *ex ante* fairness (equal opportunities, fair procedures) and *ex post* fairness (equal payoffs) to influence distributive choices (Krawczyk and Le Lec, 2010; Brock et al., 2013; Cappelen et al., 2013). We find that simple models combining *ex ante* and *ex post* fairness (e.g., Brock et al., 2013; Saito, 2013), are able to predict both the qualitative comparative-static effect of willful ignorance on the assigned punishment as well as the finding that punishment depends on consequences following willful ignorance.

There exists prior evidence that willful ignorance can be used to obtain more favorable wealth distributions, in the context of bilateral bargaining. Building on earlier experiments on bilateral bargaining with incomplete information about values, which demonstrated that more informed parties extract more favorable payoffs (Roth and Murnighan, 1982), Kagel et al. (1996) show that responders in an ultimatum game are willing to accept very unequal monetary payoffs more often when the proposer is only partly informed about the receiver's payoffs than when the proposer has complete information. Thus, a party that is ignorant about the consequences of an offer for the other party can make less favorable offers. Conrads and Irlenbusch (2013) – using a design, like ours, that is motivated by Dana et al. (2007) – confirm that this extends to (willful) ignorance: offers to another party by a proposer in an ultimatum bargaining game, who chooses to remain ignorant or cannot avoid being ignorant, are accepted more frequently than comparable offers by a fully informed proposer. Our first result that willful ignorance deflects punishment for low receiver payoffs thus concords with their finding that willful ignorance leads to higher acceptance rates of unequal proposals.

The remainder of the paper is organized as follows. Section 2 describes our experimental design. Section 3 summarizes our results with respect to the observed punishment pattern and the dictator's decisions. Section 4 discusses the predictions of different social preference models regarding the qualitative comparative-static effect of willful ignorance on punishment behavior. Finally, Section 5 concludes the paper.

2.2 Experimental Design

Our study uses one-shot binary dictator games that are modified to allow for willful ignorance and punishment. In the modified games, there are three players, as well as a move by nature that determines payoffs. Nature moves first, implementing one of two payoff states, ω_1 or ω_2, with equal probabilities, i.e., $p(\omega_1) = p(\omega_2) = 0.5$.

The state determines the relationship between a dictator's choices and the payoffs of a passive receiver, as depicted in Figure 1. More precisely, a dictator chooses between two options, a_1 and a_2. Regardless of the state, the dictator receives a payoff of 70 for choosing a_1 and 50 for choosing a_2. However, the state determines whether or not the dictator's and receiver's interests are aligned. In ω_1, the receiver receives 10 for the dictator's choice of a_1 and 50 for a choice of a_2. In ω_2 the receiver's payoffs are reversed: 50 for a choice of a_1 and 10 for a_2.

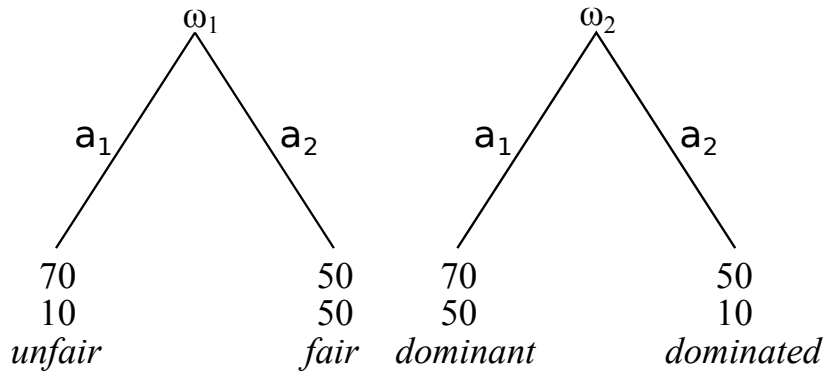


Figure 1: The Dictator's Choice Options in State ω_1 and State ω_2. The dictator's monetary payoff is shown in the top row, the receiver's payoff in the bottom row.

Figure 1 also presents labels that provide an interpretation of the dictator's actions and their consequences, conditional on the realization of a particular state. In state ω_1, a choice of a_1 leads to an *unfair* allocation in that the dictator receives the highest possible payoff and the receiver the lowest one. Conversely, a choice of a_2 in state ω_1 leads to a *fair* allocation. Thus, in ω_1 there is a conflict between what is best for the dictator and for the receiver, as in standard dictator games. However, this conflict is entirely removed in state ω_2. Here a choice of a_1 is *dominant* for a dictator who cares both about her own payoff and that of the receiver, while a_2 leads to a *dominated* allocation of 50-10.[7]

Depending on the treatment, the dictator is either informed about the realized state or not. In a *baseline* condition, the dictator is informed about the state before making a choice. In a *hidden information* condition, he is not initially informed, but he can choose whether to find out the state at no cost or remain willfully ignorant. The dictator then chooses between a_1 and a_2, either with or without knowledge of the state.

Finally, a third party can inflict punishment upon the dictator, after observing the dictator's choices (a_1 or a_2 and, in the hidden information condition, whether he remained ignorant or not), the realized state, and thus the resulting payoffs. Our primary interest is in these punishment decisions by third parties who were not directly affected by the dictator's decision.[8]

[7] We use the labels "fair," "unfair," etc. for expositional reasons in the paper. In the experimental instructions, the dictator's choice options were neutrally framed as "Option 1" and "Option 2."

[8] We use a third-party, instead of a second-party, punishment design, firstly, because we are primarily interested in broad social norms of whether willful ignorance serves as an excuse for acting in a self-interested manner and third-party punishment is often employed to study norm violations; see, e.g., Fehr and Fischbacher (2004). Moreover, measuring third-party punishment allows observing punishment assignments that are not confounded by income or

The third party has an endowment of 50 and can reduce the dictator's payoff. Punishment is costly for the third party. For each unit of own income spent by the third party, the dictator's payoff decreases by 5. Punishment is constrained in that the dictator's payoff cannot be reduced below 10. Thus, for example, if the dictator's payoff is 70 before punishment, the third party can spend any integer amount between 0 and 12 of own income, to deduct up to 60 from the dictator's payoff. If the dictator's payoff is 50, the third party can spend at most 8 units of own income, which decreases the dictator's payoff by 40.

Final payoffs are as follows. The dictator receives 70 or 50, depending on her choice of a_1 or a_2, minus the punishment assigned by the third party (five times the units of own income spent by the third party). The receiver gets either 50 or 10, depending on the dictator's decision and the relevant state. The third party's payoff is 50 minus the units of own income spent to punish the dictator.

We implemented two treatment conditions that differ only with respect to the information that the dictator possesses regarding the state.

2.2.1 Baseline

In the baseline condition, it is common knowledge that the dictator is informed about the state of the world before he makes his decision between a_1 and a_2. Thus, the dictator is fully aware of whether the choice is between the *unfair* and *fair* allocations or the *dominant* and *dominated* ones. To elicit dictator's complete strategies, we implemented the strategy method. That is, we asked each dictator how he would decide if state ω_1 were realized and how he would decide if state ω_2 were realized. Only after the dictator made both choices, he learned the actual realized state, and he knew that his choice in this state would be binding.

The third party was informed (i) about the state of the world and (ii) the dictator's choice in this realized state, and could then assign punishment to decrease the dictator's payoff. We also applied the strategy method to elicit the punishment choices. That is, we asked the third party to indicate how much she would deduct from the dictator's payoff for both possible choices by the dictator in both possible states of the world. Only after the third party made her decisions in all four possible cases, she learned the state of the world and the dictator's decision in this state. The third party knew that the chosen amount of punishment in the relevant case would be binding.

---

direct reciprocity effects. In contrast to the receiver (i.e., the second party), the third party always has an endowment of 50 points, irrespective of the resulting outcome.

2.2.2 Hidden Information

In the hidden information condition, it is common knowledge that the dictator is initially uninformed about the state of the world. Importantly, this uncertainty does not apply to the dictator's own payoffs, which are identical in both states. A choice of a_1 gives the dictator 70, while a_2 gives the dictator 50. Uncertainty thus only applies to the consequences of the two choices for the receiver's payoffs, as described in Figure 1. The dictator has the option to reveal the state before making his allocation decision. Ignorance is the default, but revealing is costless and implemented by clicking a button on the decision screen.

If the dictator remains ignorant, he will never be informed about the underlying state of the world and he will thus never learn the receiver's payoff. However, if the dictator reveals, he learns the state of the world and chooses either between the *unfair* and *fair* allocation in state ω_1, or between the *dominant* and *dominated* allocation in state ω_2.[9]

As in the baseline, we implemented the strategy method to elicit the allocation choices, where possible. That is, dictators first decided whether they wanted to acquire the payoff information or remain ignorant. If a dictator chose to remain ignorant, he then made a choice between a_1 and a_2, while if the dictator chose to acquire the payoff information, he then indicated choices of a_1 or a_2 for each of the two possible realized states. Only after the dictator made both choices, he learned the state of the world; he knew that his choice in this state would be binding.

The third party was informed of (i) whether or not the dictator revealed the state, (ii) the realized state of the world, and (iii) what choice the dictator made, either in ignorance or conditional on the realized state. The third party thus knew the state of the world even if the dictator chose to remain ignorant. We again used the strategy method to elicit the punishment decisions by third parties for all possible states and actions by the dictator. Note that there are now eight possible cases, as all four possible allocations can result either after remaining ignorant or after revealing.

[9] A basic common feature of our two treatments is that the information about the state of the world is *always* available to a decision maker, and the only difference is that willful ignorance is possible in one treatment but not in the other. This allows us to compare the consequences of a dictator's decision to remain ignorant when she could have acquired information, to situations in which the dictator is, either by default or by choice, informed. An alternative baseline, in which dictators are never informed, potentially provides insights into how judgments of punishment and blame are formed (cf. Gurdal et al., 2013), but departs from our main research question.

2.2.3 General Procedures

Before subjects entered the lab, they randomly drew a place card that specified at which computer terminal to sit and thus a subject's role and the group matching. Subjects found paper copies of the instructions at their assigned computer terminals. One third of the subjects were assigned the role of the dictator (neutrally labeled as "player A"). Two thirds of the subjects read in the instructions that they would be either in the role of the receiver ("player B") or in the role of the third party ("player C"). These subjects all made choices as third parties and they learned of their actual roles only afterward. If they were assigned the role of the third party, then the chosen amount of punishment in the relevant case would be binding. If they were assigned to the role of receiver, their decisions would have no impact on the group. This procedure enabled us to elicit punishment decisions, which are the focus of this paper, from two thirds of our subjects.[10]

We conducted four sessions of the baseline condition, with 81 subjects in total (27 subjects in the role of the dictator and 54 subjects in the role of the receiver/third party). We also conducted four sessions of the hidden information condition, with 90 subjects in total (30 subjects in the role of the dictator and 60 subjects in the role of the receiver/third party).

All sessions took place at the decision laboratory of the Department of Economics at the University of Zurich in June 2012. The experiments were computerized with the software "z-Tree" (Fischbacher, 2007) and the recruitment was conducted with the software "ORSEE" (Greiner, 2003). Subjects were students from the University of Zurich and the Swiss Federal Institute of Technology (ETH) in Zurich. Students majoring in economics or psychology were not eligible to participate. Each subject participated in only one experimental condition. Subjects' instructions included comprehension questions that had to be answered correctly before the experiment could begin. A summary of the instructions was read aloud to ensure common information regarding the instructions. An English translation of the original German instructions for the hidden information condition can be found in the online Appendix B. Sessions lasted about 50 to 60 minutes. Payoffs from the game, denominated in "points," were converted into money at the rate of 2 points to CHF 1 (about $1 at the time of the experiment) at the end of the

[10] Note that this design choice, while eliminating strategic concerns for third parties, might place third parties mentally in the role of the receivers when making their punishment decisions. Nikiforakis and Mitchell (2014) compared a punishment protocol like ours to a protocol where the role of the third party was known in advance. They found a greater demand for punishment when roles were assigned ex post but, importantly, this effect was constant across treatments and thus did not influence treatment effects.

experiment. On average, subjects earned CHF 39.80 in the baseline sessions and CHF 41.30 in the hidden information sessions. These amounts include a show-up fee of CHF 15.

2.3 Results

2.3.1 Punishment Pattern

The focus of this paper is the pattern of punishment for dictator allocation choices by third parties. Our particular interest is in studying how the dictator's choice to either remain ignorant or become informed about the receiver's payoffs influences punishment.

Figure 2 shows the average punishment that was assigned to the dictator for the different realized allocations in the baseline and in the hidden information condition.[11] The exact values can be read from Table 1. For instance, the left black bar in Figure 2 shows that the dictator receives a deduction of 19.72 points, on average, if he chooses the *unfair* allocation in state ω_1 in the baseline condition.

In accordance with prior findings on third-party punishment (e.g., Fehr and Fischbacher, 2004), the figure shows that the dictators are punished significantly more for knowingly implementing the *unfair* allocation than for the *fair* allocation. This holds true in the baseline and when the dictator chose to acquire the information in the hidden information condition. When dictators remained ignorant, the difference in punishment for implementing the *unfair* vs. *fair* allocation was smaller, but also statistically significant. Thus, regardless of the dictator's knowledge or willful ignorance of the consequence to the receiver, a choice that results in an *unfair* allocation is punished more relative to one that results in a *fair* one ($p<0.01$ in all three comparisons, using a Wilcoxon signed-rank test).[12]

[11] Averages are calculated including observations with zero punishment, i.e., we report unconditional averages.
[12] All tests reported in this paper are two-sided.

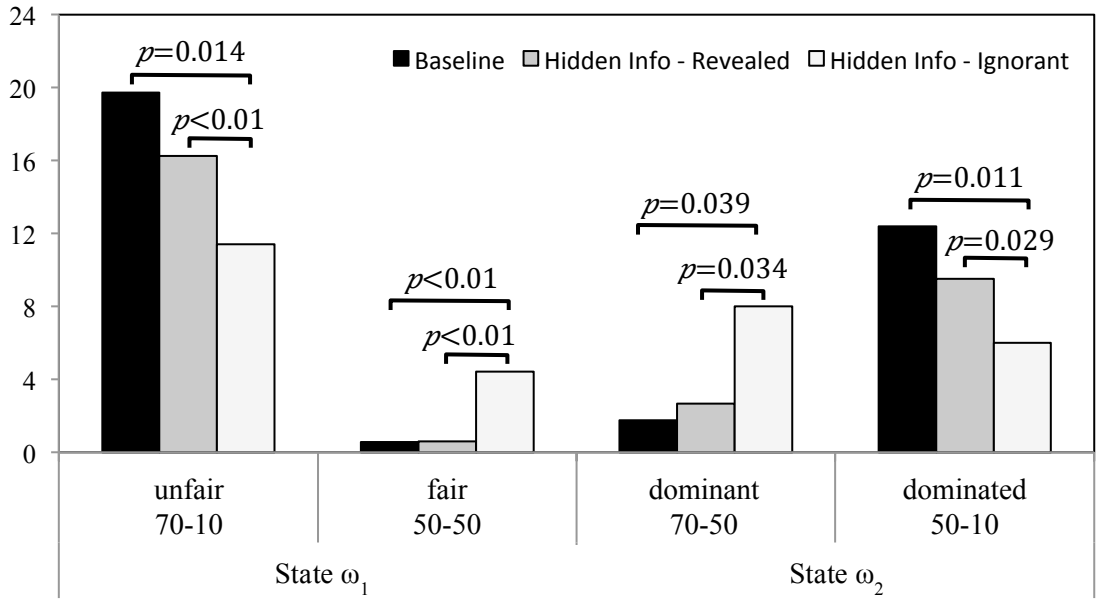


Figure 2: Average Punishment of the Dictator by the Third Party. The significance of the difference in punishment is indicated by the p-values of the respective non-parametric tests (signed-rank or rank-sum). All four comparisons between Baseline and Hidden Info – Revealed are insignificant.

Our data show, however, that willful ignorance mitigates the punishment received by a dictator whose actions result in the *unfair* allocation. A willfully ignorant dictator who chooses 70 points for himself is punished significantly less if the *unfair* allocation realizes (11.42) compared to a dictator who directly chooses the *unfair* allocation when the consequences are known – i.e., after revealing (16.25) or in the baseline condition (19.72) (Wilcoxon signed-rank test, $p<0.01$, and Wilcoxon rank-sum test, $p=0.014$, respectively).[13] Thus, our experiment reveals that willful ignorance can mitigate some of the blame and punishment received when knowingly implementing unfair outcomes.

> **Result 1**: *Willfully ignorant dictators are punished less for implementing an unfair outcome compared to dictators who knowingly chose the same outcome. Willful ignorance thus deflects blame for unfair outcomes.*

However, the opposite pattern emerges when one considers what happens in cases where the resulting allocation is the *dominant* one, which is favorable to both the dictator and the

[13] We do not find that revealing the state is treated differently from exogenously knowing the state. A comparison of the punishment for a dictator who reveals in the hidden information condition with the punishment in the baseline condition, where the dictator knows the state of the world by default, reveals no significant differences (Wilcoxon rank-sum tests, p=0.331, p=0.743, p=0.900, and p=0.196, for *unfair*, *fair*, *dominant*, and *dominated*, respectively).

receiver. Here, willfully ignorant dictators are punished significantly more (8.00) compared with dictators who choose the same *dominant* allocation after revealing (2.76) or in the baseline condition (1.76) (Wilcoxon signed-rank test, p=0.034, and Wilcoxon rank-sum test p=0.039, respectively). Thus, willful ignorance itself appears to receive blame and punishment, even when it results in an outcome favorable to everyone.

> **Result 2**: *Willfully ignorant dictators are punished more for implementing a dominant outcome compared to dictators who knowingly chose the same outcome. Willful ignorance is thus inherently blameworthy.*

Due to this opposing effect of willful ignorance on punishment, the difference in punishment between the *unfair* and the *dominant* allocation is much smaller when the dictator remained ignorant (3.42) than when he revealed the state in the hidden information condition (13.58) or in the baseline (17.96). Nevertheless, all three differences are highly significant (Wilcoxon signed-rank tests, $p<0.01$).

> **Result 3**: *Dictators, including willfully ignorant ones, are punished more if an unfair outcome is implemented than if a dominant outcome is implemented. Outcomes thus matter for punishment even under willful ignorance.*

We observe a similar pattern when a dictator chooses 50 for himself. In accordance with Result 1, if the choice is made under willful ignorance and the *dominated* allocation is implemented, the dictator is punished significantly less (6.00) compared to a dictator who chooses *dominated* after revealing (9.50) or in the baseline condition (12.41) (Wilcoxon signed-rank test, p=0.029 and Wilcoxon rank-sum test, p=0.011, respectively). However, the willfully ignorant dictator is punished significantly more if the *fair* allocation realizes (4.42) compared to a dictator who chooses *fair* after revealing (0.58) or in the baseline condition (0.56) (Wilcoxon signed-rank test, $p<0.01$, and Wilcoxon rank-sum test, $p<0.01$, respectively).[14] This finding confirms Result 2. The difference in punishment between the *fair* and the *dominated* allocation is

[14] As we report below, willfully ignorant dictators never chose 50 points for themselves. Also, none of the dictators who revealed chose *dominated* in state ω_2. In the baseline condition, only one dictator chose *dominated*. While we call the allocation (50-10) "dominated," the fact that one subject chose it highlights the possibility that it could alternatively be labeled "spiteful" or "competitive" because it maximizes the relative payoff advantage of the dictator. Punishment for a dictator who learns that the state of the world is ω_2 and nevertheless chooses (50-10) could thus be driven by third parties who want to sanction "spiteful" or "competitive" dictators. We thank a referee for suggesting this interpretation.

again much smaller when the dictator remained ignorant than when he revealed the state in the hidden information condition or in the baseline condition (1.59 vs. 8.92 and 11.85, respectively). This difference is at least marginally significant in all three cases (Wilcoxon signed-rank tests, $p=0.052$, $p<0.01$ and $p<0.01$, respectively), which is consistent with Result 3.

To summarize, we find a consistent comparative-static effect of willful ignorance on punishment. On the one hand, for resulting allocations that yield the receiver the low payoff of 10, the dictator is punished significantly *less* when he remained ignorant than when he had the payoff information (Result 1). On the other hand, for allocations that are beneficial to the receiver – i.e., when the receiver gets the high payoff of 50 – the dictator is punished significantly *more* when he remained ignorant (Result 2). Willful ignorance thus deflects blame and punishment for socially "bad" outcomes (the *unfair* or the *dominated* allocation). The fact that the dictator did not know for sure that the receiver would get a low payoff appears to serve, to some extent, as an acceptable excuse. At the same time, willful ignorance is regarded as blameworthy in itself. A willfully ignorant dictator is punished significantly more than a dictator who reveals or a dictator in the baseline condition when the receiver experiences no harm (in either the *fair* or the *dominant* allocation). Remaining ignorant means that the dictator shows some disregard for the possibility of the receiver obtaining a low payoff, and this appears sufficient for inducing punishment by third parties. Finally, we observe that outcomes matter (Result 3). Dictators always receive more punishment when their actions yield the disadvantageous outcome for the receiver, regardless of the information possessed or acquired by the dictator.

A similar pattern to the one that we observe in punishment levels also emerges when we look at the comparative-static effect of willful ignorance on the frequency of punishment, presented in Table 1. A willfully ignorant dictator who chooses a_1 and a payoff of 70 for himself is punished less often if the *unfair* allocation results (38 percent), compared to a dictator who reveals (53 percent) or to the baseline condition (61 percent) (McNemar test, $p=0.012$, and Fischer exact test, $p=0.024$, respectively). Conversely, if the *dominant* allocation results, a willfully ignorant dictator is punished more frequently (27 percent versus 13 percent, in both cases) (McNemar test, $p=0.039$, and Fischer exact test, $p=0.101$, respectively).[15] Similarly, a willfully ignorant dictator who chooses 50 for himself is punished more often if the fair

[15] In the Appendix we report the results of a hurdle model to address the question whether the effects of willful ignorance on average punishment levels are driven by different frequencies of punishment or different levels conditional on punishment taking place. The analysis suggests that differences in frequencies primarily drive our results.

allocation results and less often if the dominated allocation results, compared to a dictator who reveals or to the baseline condition, though the difference is not significant in all cases (McNemar tests, p=0.012 and p=0.180, and Fisher exact tests, p=0.010 and p=0.021, respectively).[16]

Table 1: Punishment Behavior by Experimental Condition

	Average Punishment			Frequency of Punishment		
	Baseline	Hidden Info - Revealed	Hidden Info - Ignorant	Baseline	Hidden Info - Revealed	Hidden Info - Ignorant
unfair (70-10)	19.72	16.25	11.42	0.61	0.53	0.38
fair (50-50)	0.56	0.58	4.42	0.04	0.05	0.20
dominant (70-50)	1.76	2.67	8.00	0.13	0.13	0.27
dominated (50-10)	12.41	9.50	6.00	0.50	0.37	0.28

Results 1 and 2 are further illustrated in Figure 3. The figure shows the individual third parties' punishment assignments in the hidden information condition when either the *unfair* allocation (left panel) or the *dominant* allocation (right panel) is realized. Circles above (below) the 45-degree line indicate greater (lower) punishment by third parties of dictators who revealed the state before choosing an allocation than of dictators who remained willfully ignorant. The numbers in the circles indicate the number of observations; circles without numbers represent one observation. For instance, when the *unfair* allocation realized, 27 third parties punished neither a willfully ignorant dictator nor a dictator who revealed the state of the world. Providing further support for the punishment pattern we observed earlier, of those third parties who did punish the *unfair* allocation, the majority assigned greater punishment to a dictator who revealed the state than to a willfully ignorant dictator. The pattern is reversed when the *dominant* allocation realizes: the majority of those third parties who punished assigned more punishment to a willfully ignorant dictator than to a dictator who revealed the state.[17]

[16] Consistent with our observation on levels of punishment (see footnote 13), there is no difference in the frequency of punishment between the baseline and the hidden information conditions when the dictator reveals the payoff information (Fisher exact tests, p=0.451, p=1, p=1, p=0.186 for *unfair*, *fair*, *dominant*, and *dominated*, respectively).
[17] We can also connect the behavior of individual third parties across realized allocations (i.e., across the two panels of Figure 3). Table A.2 in the online Appendix A presents the punishment patterns of individual third parties across the two outcomes and reveals that we observe similar patterns, at the individual level, that we find on aggregate.

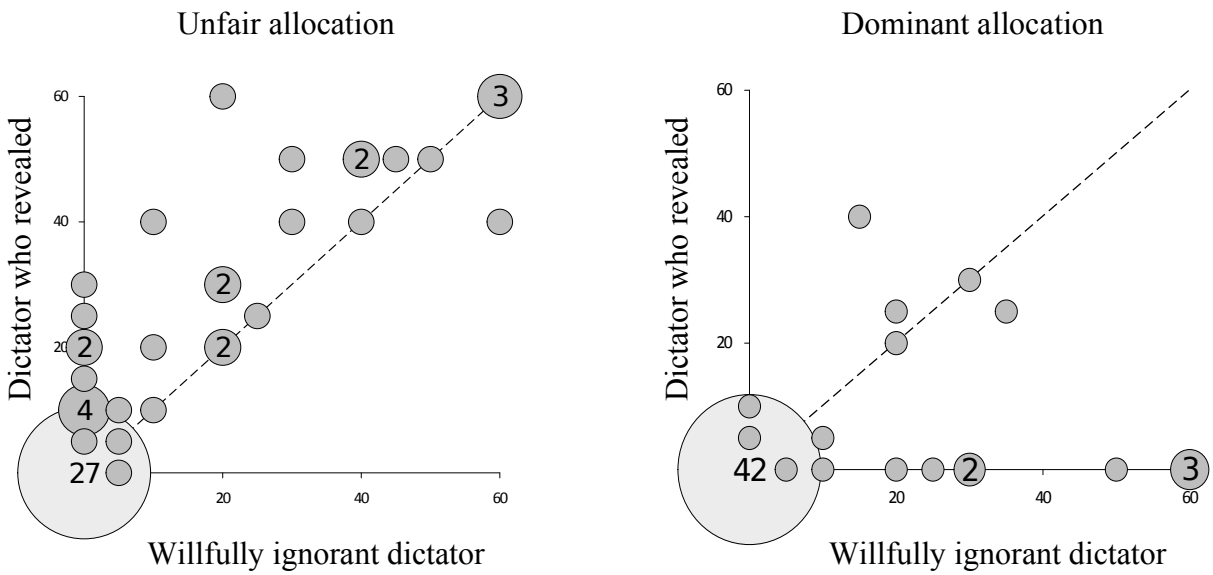


Figure 3: Individual Third Party's Punishment Assignment for the *Unfair* and *Dominant* Allocation Depending on the Dictator's Choice to either Reveal or Remain Ignorant.

2.3.2 Expected Payoffs of Dictators

We now turn to the dictators' expected payoffs for different strategies. There are four choice strategies in the baseline conditions, based on the two possible realized states and the two possible actions in each state. Because there is no uncertainty, these strategies are identified by the resulting outcomes (see Figure 1): {*unfair*, *dominant*}, {*fair*, *dominant*}, {*unfair*, *dominated*}, and {*fair*, *dominated*}. In the hidden information condition, the dictator can choose to either reveal the payoff information – in which case the same four strategies as in the baseline become available – or to remain willfully ignorant, in which case the two unconditional action choices, a_1 or a_2, are available. Table 2 shows the dictators' average expected payoffs, based on the punishment behavior of third parties, for each of these possible strategies.

Our main interest is in the effect of the dictator's choice to remain ignorant on his expected payoff. We first compare the strategies that select the same allocations. In this regard, there is little difference between the expected payoffs of a dictator who chooses to remain

ignorant and selects action a_1 (60.29) and either a dictator in the baseline (59.26) or a dictator who reveals the payoff information and selects action a_1 regardless of the realized state (60.54).[18]

Table 2: Expected Payoffs of Dictators under Different Strategies

		Baseline	Hidden Information	
			Revealed	Ignorant
{*unfair, dominant*}	$(a_1 \mid \omega_1)\ (a_1 \mid \omega_2)$	59.26	60.54	
{*fair, dominant*}	$(a_2 \mid \omega_1)\ (a_1 \mid \omega_2)$	58.84	58.38	
{*unfair, dominated*}	$(a_1 \mid \omega_1)\ (a_2 \mid \omega_2)$	43.94	47.13	
{*fair, dominated*}	$(a_2 \mid \omega_1)\ (a_2 \mid \omega_2)$	43.52	44.96	
{*unfair / dominant*}	(a_1)	-		60.29
{*fair / dominated*}	(a_2)	-		44.79

We can also compare the strategy of remaining ignorant and selecting a_1 to revealing and acting fairly in the hidden information condition or in the baseline condition (i.e., giving the receiver a payoff of 50, regardless of the state). These are the most frequently chosen strategies (see Section 3.3). While the differences are small, the expected payoff of remaining ignorant and playing a_1 (60.29) is significantly higher than the expected payoff of either of these two other strategies (58.38 and 58.84; respectively, Wilcoxon signed-rank test, $p<0.01$, and Wilcoxon rank-sum test, $p<0.01$). The observation of very small payoff differences reflects our finding that willful ignorance has two countervailing effects on punishment, described in Results 1 and 2.

2.3.3 Dictators' Strategies and Resulting Allocations

Finally, we consider the dictators' information acquisition decisions in the hidden information condition, as well as their allocation choices in both conditions.

In the baseline, 33 percent of dictators (9 of 27) chose the action a_1 regardless of the state, which corresponds to the allocations {*unfair*, *dominant*}. Almost twice as many, or 63 percent (17 of 27), chose the strategy that gave the receiver a payoff of 50 in either stage – e.g., a_2 in state ω_1 and a_1 in state ω_2, or {*fair*, *dominant*}. One subject chose action a_2 in state ω_2, implementing

[18] The difference is marginally statistically significant in the first comparison (Wilcoxon rank-sum test, $p=0.075$) but not in the second (Wilcoxon signed-rank test, $p=0.137$). For all statistical tests in this subsection, we generate a distribution of payoffs, for each strategy, using the empirical punishment behavior of the third parties.

{*fair*, *dominated*}. This overall pattern of behavior is in line with earlier results on dictator games with punishment.[19]

In the hidden information condition, 43 percent of dictators (13 of 30) remained ignorant about the consequences of their decision for the receiver.[20] All of the dictators who remained ignorant chose action a_1 {*unfair* / *dominant*}. Of those dictators who revealed the state, 12 percent (2 of 17) choose a_1 unconditionally {*unfair*, *dominant*} and 88 percent (15 of 17) choose a_2 in state ω_1 and a_1 in state ω_2 {*fair*, *dominant*}. Dictators who revealed the state thus chose the *fair* allocation in state ω_1 in the large majority of the cases, indicating that they reveal the state primarily in order to condition their allocation choice on the state of the world.

The dictators' strategies resulted in different frequencies of the possible allocations in the two conditions. In the baseline, when state ω_1 realized, 33 percent of dictators (9 of 27) chose the *unfair* allocation. The *unfair* allocation resulted with higher frequency (50 percent, or 15 of 30) in the hidden information condition. In state ω_2, the *dominant* allocation resulted almost universally in both the hidden information (30 of 30 cases) and baseline conditions (26 of 27 cases).

The fact that *unfair* allocations result more frequently under hidden information than in the baseline resembles the findings in Dana et al. (2007). In their experiment, hidden information increased the frequency of the *unfair* allocation from 26 to 63 percent. The interpretation of Dana et al. is that the possibility to remain ignorant gives subjects the moral "wiggle room" to behave self-interestedly. While similar in direction, the effect in our experiment is much smaller and not statistically significant (Fisher exact test, p=0.284). Of course, a key difference between the two experiments is the presence of a punishment stage in our design. The threat of punishment alone potentially limits the extent to which subjects are willing to act as if willful ignorance absolves them of responsibility. As we see, third parties still hold dictators responsible for their ignorance.[21]

[19] In Bartling and Fischbacher (2012), for instance, 63 percent of dictators selected a fair allocation in a binary dictator game with punishment that is comparable to our game if state ω_1 prevails.
[20] This percentage almost exactly matches the 44 percent of dictators who remained ignorant in Dana et al. (2007).
[21] Moreover, while in Dana et al., subjects who remained willfully ignorant never found out about the consequences for the receiver, dictators in our experiment received a "punishment signal" about the realized state of the world, due to the fact that third parties punished differently when the *unfair* allocation resulted than when the result was the *dominant* allocation. Thus, dictators lost some of the benefit of remaining ignorant, due to the information conveyed by punishment.

2.4 How Well do Social Preference Models Account for the Results?

In this paper, we ask the empirical question whether willful ignorance can reduce punishment for a dictator who implements an unfair allocation. Our goal was not to design an experiment to distinguish between different behavioral models of punishment and social preferences. However, it is nevertheless instructive to discuss the qualitative predictions of some leading models in the literature regarding the impact of the dictator's choice to remain willfully ignorant on the punishment by the third party. Note first that the canonical model of pure self-interest predicts no punishment at all, because it is costly. This prediction is clearly inconsistent with the data.

2.4.1 Outcome-Based Models of Social Preferences

Outcome-based models of social preferences introduce utility considerations over parties' final payoffs. For example, two leading models assume that people may dislike payoff inequalities (e.g., Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000). In the spirit of these models, suppose that the third party's punishment decisions are driven by the *ex post* payoff difference between the dictator and the receiver. Consistent with the punishment motive "ex post inequality," we observe higher punishment for allocations with higher final inequality (Result 3). For a given allocation, however, the punishment motive "*ex post* inequality" does not predict a difference based on how that allocation was produced. Our main findings (Results 1 and 2) do not support this prediction.

2.4.2 Intention-Based Models of Social Preferences

A key feature of a second class of models (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Sebald, 2010) is that players respond to the perceived intent (kind or unkind) of other players but not to realized *ex post* payoffs. The kindness of a player is typically evaluated relative to a fair "reference" payoff—e.g., the average between the highest and the lowest efficient payoff that a player can grant another player. A player's action is perceived as kind (unkind) if he believes that his action choice gives the other player more (less) than such a reference payoff. In the spirit of these models, we assume that the dictator's kindness toward the receiver drives the third parties' punishment decisions.[22]

[22] The psychological content of models of intention-based reciprocity is that unkindness triggers a reaction "in kind," i.e., punishment. While these models formally capture *bilateral* interactions, the third party in our experiment is not directly affected by the dictator's choices. Hence, our assumption that a third party's punishment decisions are driven

Consider the dictator's choice of the *unfair* allocation (70-10) in state ω_1 in the baseline or after revealing in the hidden information condition. The implementation of the *unfair* allocation is unkind because it leaves the receiver with less than the reference payoff of 30 (the average of the receiver's highest and lowest possible payoff of 50 and 10, respectively). Second, the implementation of the *fair* allocation (50-50) in state ω_1 in the baseline or after revealing in the hidden information condition is kind. Finally, remaining willfully ignorant leads to a lottery over the receiver's payoff with an expected payoff of 30, regardless of whether the dictator chooses a_1 or a_2, which is neither kind nor unkind since it corresponds precisely to the reference payoff. Qualitatively, the punishment motive "intent" thus correctly predicts Results 1 and 2. The same prediction pattern prevails in state ω_2. However, the punishment motive "intent" cannot explain Result 3. After the decision to remain ignorant, the finally resulting allocation should not influence the third party's evaluation of the dictator's kindness and thus not affect punishment.[23]

2.4.3 Models of procedural fairness

Models of procedural fairness assume people care not only about outcomes but also about the procedures that lead to these outcomes (Frey et al., 2004; Bolton et al., 2005; Trautmann, 2009; Krawczyk, 2011; Fudenberg and Levine, 2012; Brock et al., 2013; Cappelen et al., 2013; Saito, 2013). An important example of such a procedure is the notion of "equal opportunities" which can be interpreted as the idea that not only *ex post* realized payoff differences are important but also *ex ante* expected payoff differences.

Simple models of procedural fairness are suggested by, e.g., Brock et al. (2013) or Saito (2013), who extend the Fehr and Schmidt (1999) model to allow for a convex combination of *ex ante* and *ex post* payoff comparisons. Suppose the third party's punishment decisions are driven by such a convex combination of *ex ante* and *ex post* payoff differences between the dictator and the receiver and sufficient weight is placed on both ex ante and ex post payoff differences.

---

by the dictator's unkindness towards the receiver is not formally in line with these models. Third-party punishment is typically associated with norm enforcement. In that sense, one can argue that the psychological content of models of intention-based reciprocity captures the norm that one should not be unkind, and the willingness to punish violations of this norm. A similar line of argument can be made regarding our assumption that the inequality between the dictator and the receiver drives the third party's punishment in the models discussed in Sections 4.1 and 4.3.

[23] The hybrid model of outcome- and intention-based social preferences by Falk and Fischbacher (2006) makes the same qualitative prediction in our context. In their model, a player is considered as unkind if he implements an allocation that favors him in expectation. The expectation is taken at the player's decision node, so that remaining ignorant can again be treated as granting the reference payoff, as in the model by Sebald (2010).

Consider first the third party's evaluation of *ex ante* payoff differences. If the *unfair* allocation (70-10) results in state ω_1 in the baseline or after revealing in the hidden information condition, the *ex ante* payoff difference is 60. Since the *ex ante* payoff difference in case of a willfully ignorant dictator who chose a_1 is only 40, the model qualitatively predicts Result 1. If the *fair* allocation (50-50) results in state ω_1 in the baseline or after revealing in the hidden information condition, the *ex ante* payoff difference is zero. Since the *ex ante* payoff difference is 20 in case of a willfully ignorant dictator who chose a_2, the model also qualitatively predicts Result 2. The same qualitative prediction pattern prevails in state ω_2. Moreover, since *ex post* payoff differences are accounted for as well, qualitatively, the model also correctly predicts Result 3.

2.5 Conclusion

This paper studies how the opportunity to remain willfully ignorant – by avoiding information on the consequences of one's actions for others – affects the extent to which individuals are held accountable and punished by third parties for the resulting outcomes. Discussions of responsibility in political and corporate scandals are often accompanied by claims of ignorance that could have been resolved if the involved parties had sought out the relevant information. It is important, therefore, to understand whether such strategies are effective for deflecting blame and punishment.

Our findings reveal an interesting pattern. By remaining willfully ignorant, decision makers deflect some punishment when bad consequences arise, due to the fact that something good could have happened. Conversely, when good outcomes result from decisions made under willful ignorance, the fact that less desirable outcomes could have obtained provides grounds for punishment. But even under willful ignorance, punishment is still higher when bad consequences arise than when good outcomes result. Such punishment behavior by third parties is consistent with behavioral social preference models that combine *ex ante* and *ex post* fairness concerns.

For dictators in our experiment, willful ignorance is not a better strategy, in expectation, than acquiring payoff information. This is mainly because the third parties punish willful ignorance even when fortune produces a favorable outcome for the receiver. Nevertheless, the detected punishment *pattern* may have very different consequences outside the laboratory, where attention to the possibility of punishing someone is often salient only when bad outcomes arise.

In such situations, our finding that decision makers receive lighter sanctions for bad outcomes suggests that willful ignorance may be an effective strategy for circumventing blame and punishment outside the laboratory.

Interestingly, in many legal systems the "equal culpability" doctrine permits defendants who acted under willful ignorance of the existence of a fact to be treated *as if* they had possessed actual knowledge of its existence (Marcus, 1993; Husak and Callender, 1994).[24] While this observation might suggest that the law is in contradiction with people's common moral sense, as elicited in our experiment, one important difference between our experimental environment and the one governed by the legal system is that the former is a one-shot interaction while the latter is a repeated game. Deterrence of future offenses is one main function of punishment under the law, and if ignorance were a valid excuse in the law, this deterrence function would be undermined. In contrast, a deterrence motive was absent in our experimental one-shot setting.

A final aspect of our experimental design worth stressing is that information acquisition was costless for dictators. Thus, both in the baseline as well as in the hidden information condition, the relevant information was available to the dictator at no cost; the dictator merely had the opportunity to avoid seeing it in the latter condition. If information acquisition were, instead, costly, this might enhance the moral justification for remaining ignorant. For example, following the 2008 financial crisis, many individuals and institutions involved in the sale of deceptively valued and marketed investment products tried to deflect responsibility with the claim that these products were too difficult to understand, i.e., they implicitly referred to the cost of being fully informed.[25] Of course, as the cost of becoming informed increases it becomes, at some point, inefficient or even impossible for decision makers to become informed about the consequences of their actions. Hence, in some cases, ignorance may be a valid excuse for not considering the consequences of one's actions, though uncertainty and asymmetric information about these costs may complicate such considerations. These issues raise interesting questions for future research.

[24] For example, a defendant who was hired by a stranger to drive a car across the United States boarder and who claimed not to have had knowledge of the drugs that were hidden in the car was held liable to the same extent as he would have been had he had that knowledge (United States v. Jewell).

[25] See http://www.nytimes.com/2009/03/12/business/12crime.html?pagewanted=1&_r=1&th&emc=th

Bibliography

Andreoni, J., Bernheim, B. D., 2009. Social image and the 50-50 norm: A theoretical and experimental analysis of audience effects. Econometrica 77 (5), 1607-1636.

Bartling, B., Fischbacher, U., 2012. Shifting the blame: On delegation and responsibility. Review of Economic Studies 79 (1), 67-87.

Bazerman, M., Gino, F., 2012. Behavioral ethics: Toward a deeper understanding of moral judgment and dishonesty. Annual Review of Law and Social Science 8 (1), 85-104.

Bolton, G. E., Brandts, J., Ockenfels, A., 2005. Fair procedures: Evidence from games involving lotteries. The Economic Journal 115 (506), 1054-1076.

Bolton, G. E., Ockenfels, A., 2000. ERC: A theory of equity, reciprocity, and competition. American Economic Review 90 (1): 166-193.

Broberg, T., Ellingsen, T., Johannesson, M., 2007. Is generosity involuntary? Economics Letters 94 (1), 32-37.

Brock, J. M., Lange, A., Ozbay, E. Y., 2013. Dictating the risk – Experimental evidence on giving in risky environments. American Economic Review 103 (1), 415-437.

Cappelen, A. W., Konow, J., Sørensen, E. Ø., Tungodden, B., 2013. "Just luck: An experimental study of risk taking and fairness. American Economic Review 103 (4), 1398-1413.

Conrads, J., Irlenbusch, B., 2013. Strategic ignorance in ultimatum bargaining. Journal of Economic Behavior and Organization 92, 104-115.

Dana, J., Cain, D. M., Dawes, R. M., 2006. What you don't know won't hurt me: Costly (but quiet) exit in dictator games. Organizational Behavior and Human Decision Processes 100 (2), 193-201.

Dana, J., Weber, R. A., Kuang, J. X., 2007. Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. Economic Theory 33 (1), 67-80.

Dana, J., Loewenstein, G., Weber, R. A., 2012. Ethical immunity: How people violate their own moral standards without feeling they are doing so. In: De Cremer, D., Tenbrunsel, A. E. (Eds.), Behavioral business ethics: Shaping an emerging field. Psychology Press, New York.

Dufwenberg, M., Kirchsteiger, G., 2004. A theory of sequential reciprocity. Games and Economic Behavior 47 (2), 268-298.

Erkal, N., Gangadharan, L., Nikiforakis N., 2011. Relative Earnings and Giving in a Real-Effort Experiment. American Economic Review 101(7), 3330-48.

Falk, A., Fischbacher, U., 2006. A theory of reciprocity. Games and Economic Behavior 54 (2), 293-315.

Fehr, E., Schmidt, K. M., 1999. A theory of fairness, competition, and cooperation. The Quarterly Journal of Economics 114 (3), 817-868.

Fehr, E., Fischbacher, U., 2004. Third-party punishment and social norms. Evolution and Human Behavior 25 (2), 63-87.

Fischbacher, U., 2007. z-Tree: Zurich toolbox for ready-made economic experiments. Experimental Economics 10 (2), 171-178.

Frey, B. S., Benz, M., Stutzer, A., 2004. Introducing procedural utility: Not only what, but also how matters. Journal of Institutional and Theoretical Economics 160, 377-401.

Fudenberg, D., Levine, D. K., 2012. Fairness, risk preferences and independence: Impossibility theorems. Journal of Economic Behavior & Organization 81 (2), 606-612.

Greiner, B., 2003. An online recruitment system for economic experiments. In: Kremer, K., Macho, V. (Eds.), Forschung und wissenschaftliches Rechnen 2003. GWD Bericht 62, Ges. für Wiss. Datenverarbeitung, Göttingen, pp. 79-93.

Gurdal, M. Y., Miller, J. B., Rustichini, A., 2013. Why blame?. Journal of Political Economy 121 (6), 1205–1247.

Hamman, J. R., Loewenstein, G., Weber, R. A., 2010. Self-interest through delegation: An additional rationale for the principal-agent relationship. American Economic Review 100 (4), 1826–46.

Husak, D. N., Callender, C. A., 1994. Willful ignorance, knowledge, and the "equal culpability" thesis: A study of the deeper significance of the principle of legality. Wisconsin Law Review 29.

Kagel, J. H., Kim, C., Moser, D., 1996. Fairness in ultimatum games with asymmetric information and asymmetric payoffs. Games and Economic Behavior 13 (1), 100-110.

Krawczyk, M., 2011. A model of procedural and distributive fairness. Theory and Decision 70 (1), 111-128.

Krawczyk, M., Le Lec, F., 2010. ‘Give me a chance!’ An experiment in social decision under risk. Experimental Economics 13 (4), 500-511.

Lazear, E. P., Malmendier, U., Weber, R. A., 2012. Sorting in experiments with application to social preferences. American Economic Journal: Applied Economics 4 (1), 136-163.

Marcus, J. L., 1993. Model Penal Code Section 2.02(7) and willful blindness. The Yale Law Journal 102 (8), 2231-2257.

Mazar, N., Amir, O., Ariely, D., 2008. The dishonesty of honest people: A theory of self-concept maintenance. Journal of Marketing Research 45 (6), 633-644.

McDowell, A., 2003. From the help desk: Hurdle models. The Stata Journal, 3 (2), 178-184.

McGraw, K. M., 1991. Managing blame: An experimental test of the effects of political accounts. The American Political Science Review 85 (4), 1133-1157.

Nikiforakis, N., Mitchell, H., 2014. Mixing the carrots with the sticks: third party punishment and reward. Experimental Economics 17 (1), 1-23.

Ockenfels, A., Werner, P., 2012. 'Hiding behind a small cake' in a newspaper dictator game. Journal of Economic Behavior & Organization 82 (1), 82-85.

Rabin, M., 1993. Incorporating fairness into game theory and economics. American Economic Review 83 (5), 1281-1302.

Roth, A. E., Murnighan, J. K., 1982. The role of information in bargaining: An experimental study. Econometrica 50 (5), 1123-1142.

Saito, K., 2013. Social preferences under risk: Equality of opportunity versus equality of outcome. American Economic Review 103 (7), 3084-3101.

Sebald, A., 2010. Attribution and reciprocity. Games and Economic Behavior 68 (1), 339-352.

Trautmann, S. T., 2009. A tractable model of process fairness under risk. Journal of Economic Psychology 30 (5), 803-813.

Treviño, L. K., Weaver, G. R., Reynolds, S. J., 2006. Behavioral ethics in organizations: A review. Journal of Management 32 (6), 951-990.

Weaver, R. K., 1986. The politics of blame avoidance. Journal of Public Policy 6 (4), 371-398.

2.A Appendix

Regression analyses

Table A.1 reports the results of regression analyses to complement the non-parametric tests reported in the paper. Columns (1) and (2) show OLS and Tobit regressions of the punishment level on dummy variables of the dictator's decisions in the different treatments. Since the dictator's choices might affect the likelihood and amount of punishment differently, columns (3a) and (3b) report estimates from a hurdle model, an econometric specification that treats the decision to punish and the amount of punishment as two separate stochastic processes.[26] The last column reports the number of observations underlying the estimation in column (3b).

The omitted category in all regressions is the choice of the *fair* allocation in the baseline condition. The first three dummy variables measure the difference between the omitted category (*fair*) and the three other possible allocations (*unfair*, *dominant*, and *dominated*) in the baseline condition. For all three comparisons, regressions (1) and (2) show a significant and positive difference with the exception of the comparison between the fair and dominant allocations in the OLS model. The hurdle model reveals that the *unfair*, *dominant* and *dominated* allocations are punished significantly more often than the *fair* allocation. But conditional on punishment occurring, there is no significant difference in the punishment amount between the *fair* allocation and the other three allocations (the coefficients are often large in magnitude, but the large standard errors reflect the very limited punishment of the *fair* allocation in the baseline, where only two subjects chose positive punishment).

The next four dummy variables measure the difference between the baseline and the hidden information condition when the dictator reveals the state. In all four regressions, none of the four coefficients is significant, which confirms our previous finding that the punishment pattern for a dictator who reveals is the same as the pattern in the baseline condition.

Finally, the last four dummy variables measure the difference in punishment for a given allocation between a dictator who reveals and a dictator who remains ignorant in the hidden information condition. The OLS and Tobit regressions show significant differences in all four comparisons consistent with the directional results in our main analysis. The hurdle model again

[26] First, a standard probit model estimates the likelihood that a third party will punish the dictator; second, a truncated linear regression estimates the conditional likelihood of a third party punishing a certain amount (McDowell, 2003; see, also, Erkal, Gangadharan, and Nikiforakis (2011) for an example of where hurdle models are used with experimental data). The hurdle is crossed if a third party decides to punish.

reveals that these differences are driven by the frequencies of punishment, though the coefficients in regression (3b) are often large in magnitude and their sign indicates that the amount is influenced in the same direction as the decision to punish.

Table A.1: Regression Analyses

			Hurdle model		
	OLS (1)	Tobit (2)	Probability (3a)	Amount (3b)	# obs. with positive punishment
unfair (70-10)	19.17***	70.55***	0.66***	30.65	33
	(2.92)	(12.09)	(0.11)	(20.55)	
dominant (70-50)	1.20	20.85**	0.21**	-4.31	7
	(0.78)	(9.96)	(0.09)	(20.70)	
dominated (50-10)	11.85***	58.64***	0.57***	20.18	27
	(2.06)	(11.70)	(0.11)	(20.18)	
HI × unfair (70-10)	-3.47	-6.21	-0.06	-2.34	32
	(3.93)	(6.69)	(0.08)	(5.88)	
HI × dominant (70-50)	0.91	2.55	0.01	16.02	8
	(1.29)	(9.59)	(0.10)	(12.81)	
HI × dominated (50-10)	-2.91	-8.80	-0.11	1.69	22
	(2.83)	(6.84)	(0.07)	(5.65)	
HI × fair (50-50)	0.03	3.93	0.05	-11.16	3
	(0.61)	(13.86)	(0.13)	(24.70)	
HI × ignorant × unfair (70-10)	-4.83***	-11.20***	-0.12***	-0.92	23
	(1.26)	(2.88)	(0.04)	(4.29)	
HI × ignorant × dominant (70-50)	5.33**	19.51**	0.16**	15.97	16
	(2.10)	(7.75)	(0.07)	(10.66)	
HI × ignorant × dominated (50-10)	-3.50***	-9.15**	-0.07*	-7.90	17
	(1.29)	(3.88)	(0.04)	(5.05)	
HI × ignorant × fair (50-50)	3.83***	28.82***	0.26***	26.77	12
	(1.38)	(9.84)	(0.09)	(18.68)	
Constant	0.56	-59.92***	-	-0.78	2
	(0.47)	(12.16)		(21.39)	
Observations	696	696	696	202	202
(Pseudo) R^2	0.16	0.05	0.14	0.11	

Notes: The dependent variable in regression (1), (2) and (3b) is the size of the punishment reduction received by a dictator. The dependent variable in regression (3a) is a dummy that equals 1 if the third party punishes. "Probability" reports the marginal effects from a probit regression calculated at the mean. "Amount" is a linear regression truncated at 0. The omitted category in all regressions is the choice of the fair allocation in the baseline condition. "HI" indicates the hidden information condition. Robust standard errors clustering at the subject level are reported in parentheses.
*** denote significance at 1 percent, ** at 5 percent, and * at 10 percent.

Chapter 3

A Theory of Causal Responsibility Attribution[1]

"No snowflake in an avalanche ever feels responsible." Stanislaw Jerzy Lec

3.1 Introduction

Perceptions of responsibility oftentimes guide the attribution of blame and praise. When we like an event, we typically praise the person responsible for its implementation, and when we don't like an event, we blame the person responsible for its implementation. Through this channel, responsibility perceptions play a crucial role in many environments that are of traditional interest to economists. For example, in labor markets, workers and managers are fired or rewarded depending on their responsibility for the failure or success of a project. In political economy contexts, people make voting decisions based on the attribution of responsibility to politicians for the implementation of reforms and economic outcomes. In behavioral ethics, the question arises whether firms or their customers are responsible for negative externalities of the production process. Ultimately, the design of institutions like hierarchies and voting rules can be guided by judgments of

[1]This paper must be cited as: Engl, F. (2015): "A Theory of Causal Responsibility Attribution," Working Paper. I want to thank my supervisor, Roberto Weber, for excellent guidance throughout the project, and my co-supervisor, María Sáez-Martí, for numerous helpful comments and discussions. I also thank Björn Bartling, Ernesto Dal Bo, Pedro Dal Bo, Lea Cassar, Guillaume Fréchette, Lea Heursen, Matthew Jackson, Dorothea Kübler, Igor Letina, Nick Netzer, Arno Riedl, Andrew Schotter, Eldar Shafir, Dirk Sliwka, Joel Sobel and Ran Spiegler and seminar and conference participants at the University of Cologne, the University of Munich, the Zurich Workshop in Economics, the ZWEBER 2014, the 2015 Zurich-Basel Workshop in Micro Theory, the NYU-CESS 8th Annual Experimental Political Science Conference, and the 2015 Morality, Incentives and Unethical Behavior Conference at UC San Diego for helpful discussions and comments.

when it is better to diffuse or to focus responsibility on individual agents. As these examples demonstrate, responsibility perceptions become especially important when events are implemented through the interaction of several parties and potentially nature, i.e. when the question arises who is more or less responsible for an event. Understanding the determinants of responsibility perceptions in group settings and their implications for the attribution of blame and praise are therefore important research questions in economics.

Despite their importance, little research in economics has dealt with the evaluation of responsibility and an established theoretic notion of responsibility does not exist.[2] In this paper, I study the attribution of responsibility to agents for the implementation of an event, when the event's implementation depends on the interaction of multiple agents and/or nature. In particular, I develop a notion of *causal responsibility*, which is based on a counterfactual-reasoning approach, incorporate the notion into a framework of *responsibility preferences* and study its implication for the allocation of punishment and reward in a two-stage game. Responsibility preferences imply that, in addition to their preferences over monetary payoffs, agents have a taste to reward (punish) other agents for the implementation of what they judge as good- or bad events, to the extent that those agents are causally responsible for the event.[3] I further demonstrate that the predictions of the theory can explain existing evidence from experiments that prominent existing theories have difficulties explaining.

The notion of *causal responsibility* attempts to objectively capture the causal importance of an agent's action for the implementation of an event.[4] It is especially important, because it often serves as a necessary condition for other factors that play a role for the attribution of punishment and reward. For example, Thompson (1980), in his work on the responsibility of public officials, writes that, "unless an official's action is at least a causal factor of an event, it is hard to see why the question should arise of holding that official, rather than anyone or everyone else, responsible for it." In the realm of moral judgments,

[2] A growing experimental literature studies the implications of responsibility perceptions for the attribution of punishment (Bartling, Fischbacher, and Schudy, forthcoming; Duch, Przepiorka, and Stevenson, 2014). A notable theoretical exception, that will be discussed in greater detail later on, is the experimental study by Bartling and Fischbacher (2012), who develop a notion of responsibility that measures an agent's responsibility for an event by his impact on the probability that the event is implemented compared to some reference probability. In the theoretical work of Prendergast (1995) and Sliwka (2006), responsibility is allocated ex ante by a principal to a worker for the implementation of a task and Manove (1997) models "responsible jobs" as those in which a worker's effort can influence the output.

[3] A large number of experimental studies has shown that people are willing to incur costs in order to punish and reward other people (Fehr and Gachter, 2002; Boyd, Gintis, Bowles, and Richerson, 2003; Nikiforakis and Mitchell, 2013) and that this holds even for unaffected third parties (Fehr and Fischbacher, 2004; Leibbrandt and López-Pérez, 2012; Bartling, Engl, and Weber, 2014). I am therefore taking the willingness to punish and reward as given and study the comparative-static effects of changes in causal responsibility on punishment and reward.

[4] Of course, other approaches to responsibility exist. For example, Hart (1968) categorizes four different notions of responsibility that play a role in legal contexts: role-responsibility, causal-responsibility, liability-responsibility, and capacity-responsibility.

the psychologists Darley and Shultz (1990) write that "judgments of moral responsibility presuppose those of causation. If the protagonist is judged not to have caused the harm, then there is no need to consider whether he is morally responsible for it," and, going even further, Sloman, Fernbach, and Ewing (2009) argue that the "causal structure is so central to moral judgment that representations of causal structure, causal models, serve as the representational medium for appraising and reasoning about the morality of events." Thus, when ignoring the underlying causal structure, predictions for the attribution of punishment and reward can be faulty.

Understanding the link between perceptions of causality and responsibility has therefore long been recognized as an important area of research in many social sciences. For example, in law, Hart (1968) and Hart and Honore (1985) argue that causality is the prime determinant of responsibility which, in turn, determines legal liability. Wright (1985, 1988) and Moore (2009) discuss specific models of causation and their relationship with responsibility and legal liability. In political science, causality-based models of responsibility are used as a justification for why people should vote in elections (Goldman, 1999) and as an explanation for the difficulty of apportioning appropriate blame and praise to public officials (Thompson, 1980). Furthermore, Gomez and Wilson (2003) study the claim that more sophisticated voters should be better at understanding causal mechanisms and therefore attribute more responsibility for economic outcomes to the party that rules congress as opposed to the party of the president. Abramowitz, Lanoue, and Ramesh (1988) and Iyengar (1996) study how personal finances and the media, respectively, can influence perceptions of the government's causal responsibility. Philosophy has considered the dilemma of how to ration medicine when it is in limited supply and suggested that those who are causally responsible for their illness, for example, by smoking, should be given lower priority (Dietrich, 2002). In addition, several studies discuss the challenges of causal and moral responsibility concepts from a philosophical point of view (Bunzl, 1979; Sober, 1988; Miller, 2001). In economics, Berg (1982) argues that a consumer's causal responsibility for increases in supply capacity should be considered when designing peak-load pricing schemes in electricity markets. In psychology, evaluations of causality have long been an important part in the judgment of responsibility and blame, both in the tradition of attribution theory (Heider, 1958; Shaver, 1985) as well as in culpable control theory (Alicke, 1992, 2000; Alicke, Buckingham, Zell, and Davis, 2008), and are generally seen as a prerequisite for the attribution of blame (Darley and Shultz, 1990; Schlenker, Britt, Pennington, Murphy, and Doherty, 1994; Weiner, 1995; Sloman, Fernbach, and Ewing, 2009; Malle, Guglielmo, and Monroe, 2014). The role of perceptions of causality for the attribution of blame has been confirmed in many experimental psychological studies (Spellman, 1997; Lagnado and Channon, 2008; Cushman, 2008).

However, many of these studies rely on simple heuristics for the evaluation of causal

responsibility, oftentimes boiling down to a simple Yes/No decision. Hence, on first sight, it might look like the study of causal responsibility doesn't require much insight and can therefore safely be ignored in economic models. And, indeed, when a single agent chooses an action that directly translates into an event, individual causal responsibility is straightforwardly established as only that agent and his choice of action determined the event. However, when events are implemented through the interaction of several parties and/or nature, determining each parties degree of causal responsibility for the event is not trivial. For example, when a single research group accomplishes an important scientific breakthrough, like finding a cure for cancer, that research group is attributed full responsibility for the outcome and therefore receives praise. However, what if two research groups simultaneously and independently find a cure for cancer? Are they both perceived as a cause for the outcome and thus both praised as much as the single group, or not at all, because no group alone was pivotal for the breakthrough? How does the answer change if there are ten or more successful research groups?

Intuition suggests that two successful research groups are still held responsible, and thus praised, to some degree, but not as much as the single successful research group. In order to be able to handle such questions theoretically, I extend and bring into the economic framework a notion of responsibility that was pioneered by Chockler and Halpern (2004) in the artificial intelligence literature and that is based on the *structural account* of causation (Pearl, 2000; Halpern and Pearl, 2005; Woodward, 2003).[5] In Chockler and Halpern's notion, the responsibility of A for the realized event B inversely depends on the minimum number of changes that have to be made to the specific context in order to make B counterfactually depend on A. Hence, in the example above, two successful research groups are less responsible for the scientific breakthrough than the single successful research group. But they are more responsible than if 10 groups would have independently achieved the same breakthrough. Recent experiments in psychology that elicit non-incentivized responsibility ratings mostly confirm the comparative-statics predictions of such a counterfactual-reasoning-based responsibility notion (Gerstenberg and Lagnado, 2010; Zultan, Gerstenberg, and Lagnado, 2012; Lagnado, Gerstenberg, and Zultan, 2013). For example, in Lagnado, Gerstenberg, and Zultan (2013) subjects were told that for a hypothetical team of four, consisting of members A, B, C and D, to be success-

[5]Put simply, under the structural account, A is a cause for B, if there exist hypothetical contingencies in which B counterfactually depends on A. In the case of two successful research groups, even though the outcome - having a cure for cancer - does not counterfactually depend on any of the two in the actual realization, there exists a possible hypothetical contingency, namely the one in which only one found a cure and the other didn't, in which the event counterfactually depends on the successful group. Thus, the structural account allows both research groups to be a cause of the event. Such counterfactual reasoning is not only used as a tool in the theoretical causality literature, but has also been shown by psychologist to be a common method that people employ when assessing causality (Kahneman and Tversky, 1982; Kahneman and Varey, 1990; Roese, 1997; Spellman, 1997; Spellman and Mandel, 1999; Alicke, 2000). For an extensive recent summary of different approaches to causality, see Beebee, Hitchcock, and Menzies (2012).

ful, both A and B as well as one out of C and D have to succeed in their individual task. Subjects were then asked to rate the responsibility of A for the failure of the team task, if A failed and varying combinations of B, C and D also failed. Subject attributed higher responsibility to A, the smaller the number of changes that had to be made to the other members' outcomes, in order to make the team's failure counterfactually depend on A's failure.

While lending support to a notion of causal responsibility that is based on counterfactual reasoning, these studies don't show how responsibility perceptions influence actual behavior. However, psychological factors like responsibility perceptions are typically only of interest to economists, if they induce behavioral consequences, such as changes in choice-behavior. In this study, I, therefore, firstly formalize an extended notion of causal responsibility in game-theoretic notation, making it tractable for economic modeling, and, secondly, incorporate causal responsibility in a preference framework that allows to study its behavioral implications for the allocation of punishment and reward.[6] Furthermore, I study how those subject to responsibility-driven attribution of punishment and reward react to it, thereby analyzing the consequences of responsibility perceptions for equilibrium outcomes. I show how, depending on the specific environment, causal responsibility perceptions can induce distinctively different equilibrium outcomes in which, in some cases, causal responsibility for an event is maximally diffused between all, and, in other cases, maximally focused on some of the involved agents. Finally, I test the predictive power of the notion causal responsibility for the allocation of punishment in data from existing, incentivized experiments. I find that it can explain observed punishment patterns in many cases more successfully than existing theories and that it remains a highly significant predictor for punishment even after controlling for several other potential punishment motives.

My overall notion of causal responsibility features a convex combination of an ex ante and an ex post component. Ex post causal responsibility measures how causally responsible the action of an agent turned out to be for the implementation of an event, taking the realized actions of all agents and (potentially) nature into account. It crucially depends on the "distance" of an agent's action from being pivotal for the event, where "distance" is measured by the number of hypothetical changes to the realized actions of the other agents (and nature) that it takes to make the action of the agent under consideration pivotal for the event. Therefore, an agent is said to bear *full ex post causal responsibility* for an event, if his action is pivotal for it. A non-pivotal agent's degree of *partial ex post*

[6] To highlight the importance of such formalization, one might draw a parallel to the theory of inequity aversion (Fehr and Schmidt, 1999). While it was well known how to measure inequity and that people might care about inequity, only its formulation in terms of notation tractable for economic modeling and its inclusion in a preference framework facilitated the study of its impact on behavior and thus its implications for economic contexts, such as markets, wage setting, etc.

causal responsibility for an event inversely depends on the necessary number of changes to the actions of the other agents (and nature) to make the agent under consideration pivotal for the event. A non-pivotal agent, who, given his action, could never be pivotal for an event bears *no ex post causal responsibility* for the event. The ex ante causal responsibility component captures that, in the presence of nature, there can exist objective uncertainty about the degree of ex post causal responsibility that follows from a given action. Therefore, agents whose actions induce a higher expected level of ex post causal responsibility for an event are attributed higher ex ante causal responsibility for the event. The convex combination of ex post and ex ante causal responsibility can be interpreted as follows: Placing weight on ex ante causal responsibility only can be understood as deontological responsibility attribution, whereas placing weight on ex post causal responsibility only can be understood as consequentialist responsibility attribution. In most real-world cases, agents are expected to place weight on both, deontological and consequentalist motives.[7]

The model consists of a two-stage game with complete information. In the first stage, several agents with standard preferences and potentially nature simultaneously choose actions. The realized actions collectively implement a stage-1 event. In the second stage, another agent, with "responsibility preferences", evaluates the possible stage-1 events and the stage-1 agents' causal responsibility for them. He then has the opportunity to punish or reward the stage-1 agents through an allocation decision.[8] Intuitively, the stage-1 agents could be thought of as firms which decide whether to operate with a clean or a dirty technology. The dirty technology is cheaper, but, if too many firms use it, it leads to the destruction of the environment. Firms only intend to maximize their profits and therefore don't care about the destruction of the environment. However, another agent, who can be thought of as a consumer, cares about the destruction of the environment and evaluates it as a bad event. He observes the production choices of the firms and will, when making his consumption decision, take their causal responsibility for the destruction of the environment into account. In equilibrium, the firms rationally anticipate the effects of their actions on the reaction of the consumer and adopt their behavior accordingly. The existence of a subgame perfect Nash equilibrium in which the environment is destroyed depends on the number of firms that are operating in the market. If enough firms are present, there exists an equilibrium in which all firms use the dirty technology and the environment is destroyed. In this case causal responsibility is diffused enough among

[7] In comparison with Chockler and Halpern's notion of responsibility, my notion differs in two important aspects: First, I allow that ex post causal responsibility is also evaluated for hypothetical, unrealized events, making it possible to assign responsibility for events that did not happen, but could have happened, given an agent's action. Second, I allow objective uncertainty to play a role through its impact on ex ante causal responsibility.

[8] Throughout, the paper adopts a positive approach to responsibility assessments, i.e. it seek to understand and model how responsibility perceptions influence how people actually reward and punish instead of studying how they should reward and punish.

them such that the consumer's negative reaction is outweighed by the fixed gain of lower production costs.

In economics, the only other paper that I am aware of which provides a theoretic notion of ex post responsibility allocation in a multi-agent context is the experimental study of Bartling and Fischbacher (2012). Their measure attributes most responsibility for an event to the agent whose action led to the largest increase in the probability that the event is implemented compared to the ex ante belief of the agent evaluating responsibility, which is assumed to be based on his belief about the average play in the game. My approach differs in several important aspects. First, their notion is not included in a preference framework or a game-theoretic equilibrium concept. Hence, it doesn't model how responsibility perceptions enter the utility function and how such preferences might play out in equilibrium. Second, it crucially depends on the beliefs of the agent who evaluates responsibility. Those beliefs are assumed to concord with "average play". If average play is that an action is taken with certainty, as in a pure-strategy equilibrium, no agent who takes that action is attributed any responsibility for the event, even if his action is pivotal. Third, given the construction of their measure, the sum of responsibility cannot exceed one, i.e. it is not possible to have multiple agents who are each fully responsible for an event. Two agents who each increased the probability of an event happening to the same extent each have responsibility 1/2, independent of whether they are both pivotal, or not. Therefore, their measure is a diffusion of responsibility measure (Darley and Latané, 1968). Fourth, moves of nature are ignored, while my measure specifically recognizes the role of nature in the formation of causal responsibility perceptions.

The theory of causal responsibility can also be compared to social preference theories, some of which are able to make predictions for punishment and reward in scenarios as the ones described above.[9] For example, outcome-based social preference theories predict that people engage in punishment and reward, when it can be used to decreases final payoff inequalities (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000). These theories therefore predict punishment and reward independent of the responsibility of the agents. Intention-based social preference theories, on the other hand, predict punishment and reward as a way to reciprocate unkind with unkind and kind with kind behavior (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004). Players evaluate kindness by comparing the

[9]One concept that is related at first sight is Konow's concept of distributional justice, which is based on a notion of accountability (Konow, 1996, 2000, 2001). Konow asks the question of how to justly allocate an endowment among the agents that produced it. He concludes that allocations should be driven by the variables that the agents could control, but not by those they couldn't control. For example, when agents' production depends on their effort but also on their innate ability, a just distribution would be proportional to the agents' efforts, but not to their innate ability. The main distinguishing feature compared to my concept is that I am not interested in finding a "just" distribution, but on how causal reasoning influences actual behavior. While higher effort *can* correlate with a higher degree of causal responsibility, this is not necessarily the case.

payoff that was given to them to a reference payoff, often assumed to be the average between the highest and the lowest efficient payoff possible. Crucially, kindness evaluations presuppose full causal responsibility, as only when one is pivotal could one have any influence on the final payoff. Therefore these theories are unable to make sharp predictions for punishment and reward when pivotality is not achieved, which oftentimes happens in the type of group situations that I am most interested in.[10]

The remainder of the paper is organized as follows: Section 2 introduces the notion of causal responsibility. Section 3 incorporates causal responsibility in a framework of responsibility preferences and analyzes a simple example in order to distinguish the prediction from other prominent theories. In Section 4, existing experimental evidence is analyzed in light of the theory of responsibility preferences. Section 5 discusses some caveats and lays out avenues for future research and Section 6 concludes.

3.2 The notion of causal responsibility

Before defining *responsibility preferences* and the game structure per se, this section formalizes the notion of causal responsibility in a simultaneous-move environment. The framework is as follows: There is a finite set I of agents indexed by $i \in I = \{0, 1, ..., n\}$ with nature being indexed as agent 0. Each agent i takes an action $a_i \in A_i$, where A_i is the finite set of feasible actions of agent i. The set of feasible action profiles is denoted as $A = \prod_{i \in I} A_i$. Moves of nature follow an exogenously given probability distribution over nature's action space A_0, denoted by ω.

The *realized* action profile $\mathbf{a} = (a_0, ..., a_n)$ determines the event $x \in X$, where X is the set of all possible events. The function f relates actions to events $f : A \to X$. Oftentimes, events are binary: a reform is either implemented, or not; a team project is either successful, or not; climate change happens, or not. Similarly, actions are often binary choices: a senator votes for, or against a reform; effort is provided, or not; a firm decides to use dirty, or clean technology.

In the following, I firstly provide a measure of *ex post causal responsibility*, which is based on the realized actions profile. Then, I provide a measure of *ex ante causal responsibility*, which is based on nature's probability distribution and the chosen actions. *Overall causal responsibility* is finally defined as a convex combination of ex ante and ex post causal responsibility.

[10] From a philosophical perspective, another way to distinguish intentionality from causal responsibility is to ask whether the concept could also be applied to purely physical objects instead of people. While it is natural to say an avalanche was causally responsible for the destruction of a village, it is not natural to say that the avalanche's intention was unkind.

3.2.1 Ex post causal responsibility

An agent i's ex post causal responsibility for an event x, captured by the function $r^{EP}_{i,x}$, is evaluated after the *realized* action profile $\mathbf{a} = (a_0, ..., a_n)$ and the accompanying *realized* event $f(\mathbf{a})$ are observed. I model ex post causal responsibility after the intuition of Chockler and Halpern (2004) who introduce a responsibility function in which responsibility is inversely related to the minimum number of changes that have to be made in a specific context in order to achieve pivotality. Therefore, I first introduce a function that measures the "distance" that an agent's action is away from being pivotal for the realization of the event x, given the *realized* action profile $\mathbf{a}$. This "distance" function depends on the number of changes one has to make to the actions of the other agents and nature in order to make the action of the agent under consideration pivotal for the event. The "distance" measure is then used to construct the causal responsibility function.

In order to formally measure the "distance" of agent i's action from being pivotal for event x, I first define a set $\tilde{A}(x, a_i)$. This set comprises all action profiles $\tilde{\mathbf{a}} \in A$ for which it holds that, if the other agents (including nature) play actions $\tilde{\mathbf{a}}_{-i}$, then agent i's action a_i is pivotal for event x. Being pivotal means that there exists an alternative action for agent i, $\tilde{a}_i$, such that, if agent i would switch from a_i to $\tilde{a}_i$, event x would not be implemented anymore. Formally, $\tilde{A}(x, a_i) = \{\tilde{\mathbf{a}} \in A \mid f(a_i, \tilde{\mathbf{a}}_{-i}) = x \text{ and } f(\tilde{a}_i, \tilde{\mathbf{a}}_{-i}) \neq x\}$. Next, I let $c(\mathbf{y}, \mathbf{z})$ be a function that counts the number of different entries in vectors $\mathbf{y}$ and $\mathbf{z}$. Hence, $c(\tilde{\mathbf{a}}, \mathbf{a})$ yields the number of actions in which the realized action profile $\mathbf{a}$ differs from an alternative action profile $\tilde{\mathbf{a}}$.

Definition 1. *The distance of agent i's action a_i from being pivotal for event x, $d_{i,x}$, is defined as*

$$d_{i,x}(\mathbf{a}) = \begin{cases} \min_{\tilde{\mathbf{a}} \in \tilde{A}(x,a_i)} c(\tilde{\mathbf{a}}, \mathbf{a}) & \text{if} \quad \tilde{A}(x, a_i) \neq \emptyset \\ \infty & \text{if} \quad \tilde{A}(x, a_i) = \emptyset. \end{cases} \tag{3.1}$$

Agent i's distance from pivotality is defined to be infinite, if there exist no changes to the other agents' actions (including nature) that would make agent i's action pivotal for the event x. On the other hand, if there exist changes that would make his action pivotal, then distance is defined as the minimum number of changes necessary in order to do so. The minimum is bounded below by 1 because the change of action, from a_i to $\tilde{a}_i$, that agent i has to make in order to change the event is accounted for as well. It is also bounded above by $|I|$, the cardinality of set I. Hence, $d_{i,x} : A \to \{1, 2, ..., |I|, \infty\}$.[11]

[11] The distance function, in this form, treats every change of action identically. I discuss potential extensions to alternative formulations in section 3.5.

The distance function is used to formulate an axiom, which relates distance from pivotality to the ex post causal responsibility function, $r_{i,x}^{EP}$.

Axiom 1 (Monotonicity)**.** *For any two agents $i \neq j \in I$, $r_{i,x}^{EP}(\mathbf{a}) \geq r_{j,x}^{EP}(\mathbf{a})$ if and only if $d_{i,x}(\mathbf{a}) \leq d_{j,x}(\mathbf{a})$.*

The monotonicity axiom states that for any two agents, the agent with smaller distance from pivotality for event x has a higher degree of ex post causal responsibility for event x. Therefore, in the example from the introduction, two independently successful research groups each bear less ex post causal responsibility for a scientific breakthrough than a single successful research group. But they bear more ex post causal responsibility than each of ten independently successful research groups. Any ex post causal responsibility function that is decreasing in distance would satisfy this axiom. In order to increase tractability, I define ex post causal responsibility to have a specific form, which has some desirable properties in addition to satisfying the monotonicity axiom.

Definition 2. *An agent i's degree of ex post causal responsibility for event $x \in X$, $r_{i,x}^{EP}$, is defined as*

$$r_{i,x}^{EP}(\mathbf{a}) = \frac{1}{d_{i,x}(\mathbf{a})}. \tag{3.2}$$

Since distance can only take on discrete values, the ex post causal responsibility function is discrete, $r_{i,x}^{EP} : A \to \{0, \frac{1}{|I|}, \frac{1}{|I|-1}, ..., 1\}$. Ex post causal responsibility values, however, are easy to calculate and interpret. In the case of the research groups, one successful research group has ex post causal responsibility of 1 for the event, two groups have ex post causal responsibility of $\frac{1}{2}$, three of $\frac{1}{3}$, and so forth. Furthermore, the function possesses two desirable properties, which, together, are unique to this function.[12]

First, since the distance, $d_{i,x}(\mathbf{a})$, is bounded below by 1 (in this case, agent i is pivotal for event x without any changes to the others' actions), the monotonicity axiom leads to the corollary that the ex post causal responsibility function, $r_{i,x}^{EP}$, achieves a maximum if $d_{i,x}(\mathbf{a}) = 1$. Furthermore, as the largest value $d_{i,x}(\mathbf{a})$ can achieve is infinity, the ex post causal responsibility function, $r_{i,x}^{EP}$, achieves a minimum if $d_{i,x}(\mathbf{a}) = \infty$. The ex post causal responsibility function reflects these bounds and lets the lower and the upper bound of ex post causal responsibility be 0 and 1. The scale from 0 to 1 has the advantage that it is immediately interpretable as *no* ($r_{i,x}^{EP}(\mathbf{a}) = 0$), *partial* ($r_{i,x}^{EP}(\mathbf{a}) \in (0, 1)$), and *full* ($r_{i,x}^{EP}(\mathbf{a}) = 1$) ex post causal responsibility.

[12]The function is therefore characterized by properties P.1 and P.2 (for a theorem and the proof, see Appendix 3.A.1).

Property 1 (Boundedness)**.** *For any agent* $i \in I$*,* $r^{EP}_{i,x}(\mathbf{a}) = 0$ *if and only if* $d_{i,x}(\mathbf{a}) = \infty$ *and* $r^{EP}_{i,x}(\mathbf{a}) = 1$ *if and only if* $d_{i,x}(\mathbf{a}) = 1$*.*

Second, the function also satisfies a proportionality property, which governs the relationship between causal responsibility and distance from pivotality in between the bounds of zero and one. It says that the proportion of ex post causal responsibility of any two agents for an event x is inversely related to the proportion of their distance from pivotality for that event. Hence, an agent whose distance from pivotality for an event is twice as big as another agent's, ceteris paribus, has half the ex post causal responsibility for that event. Applied to the example, two independently successful research groups have half the ex post causal responsibility for a scientific breakthrough than a single successful research group and three independently successful groups have three times as much ex post causal responsibility than nine independently successful groups.

Property 2 (Proportionality)**.** *For any two agents* $i \neq j \in I$*,* $\frac{r^{EP}_{i,x}(\mathbf{a})}{r^{EP}_{j,x}(\mathbf{a})} = \frac{d_{j,x}(\mathbf{a})}{d_{i,x}(\mathbf{a})}$*.*

Of course, other possibilities than a fixed relative relationship between distance and ex post causal responsibility are conceivable. For example, the effect of distance on causal responsibility could be decaying with distance such that doubling the distance halves causal responsibility when distance is small but does not do so when distance is large. Whether this fixed relationship holds is therefore an empirical question.

Importantly, the event under consideration, x, need not be equal to the realized event $f(\mathbf{a})$. Hence, the framework allows to evaluate ex post causal responsibility even for hypothetical, i.e. unrealized events. Take the example of a person who decides about shooting or not shooting a gun. If he shoots, he has a 50 percent probability of killing another person. If he doesn't shoot, nothing happens. Suppose he shoots and kills the other person. In that case, he is pivotal and thus fully ex post causally responsible for the killing. However, it is also natural to assign responsibility for events that could have happened had nature chosen otherwise. Suppose he shoots, nature intervenes, and he does not kill the other person. In that case, we would still assign partial ex post causal responsibility to the person for the hypothetical event of killing the other person, for which the person would have been pivotal had nature chosen differently.

To illustrate the difference between partial and no ex post causal responsibility, consider the example of a team of workers that has the task to complete a project successfully. If one worker decides to shirk, but his coworkers work hard and successfully finish the project, the shirking worker has no ex post causal responsibility for the success of the project, because, for no alternative actions of his coworkers, would his shirking be pivotal for the success. On the other hand, he is partially ex post causally responsible for the

hypothetical event that the project failed, because, if just enough other workers would have also decided to shirk such that the project fails, his shirking would have been pivotal for the failure.

Example

The following example demonstrates how ex post causal responsibility is evaluated in two simple environments. In both environments, agent 1 can take actions u and d and agent 2 can take actions l and r.

The two environments are shown in Table 3.1. In the *substitutes* environment, event e_1 is implemented whenever agent 1 chooses u or agent 2 chooses l, otherwise event e_2 is implemented. In the *complements* environment, event e_1 is implemented whenever agent 1 chooses u and and agent 2 chooses l, otherwise event e_2 is implemented.

	l	r
u	e_1	e_1
d	e_1	e_2

Substitutes environment

	l	r
u	e_1	e_2
d	e_2	e_2

Complements environment

Table 3.1: Two simple environments.

In the following, the ex post causal responsibility of agent 1 for event e_1 is evaluated.[13] The resulting ex post causal responsibility values are shown in Table 3.2.

In the *substitutes* environment, when agent 1 chooses u and agent 2 chooses r, i.e. the realized action profile is (u, r), agent 1 is pivotal for event e_1 and, thus, his ex post causal responsibility for event e_1 is $r^{EP}_{1,e_1}(u, r) = 1$. In this case, agent 1 is said to bear *full* ex post causal responsibility for event e_1. If, on the other hand, the action profile (u, l) realizes, then agent 1 is not pivotal for event e_1. However, he would be, if agent 2 would change his action. Thus, the distance from pivotality for agent 1 and event e_1 is $d_{1,e_1}(u, l) = 2$ and $r^{EP}_{1,e_1}(u, l) = \frac{1}{2}$. Then, agent 1 is said to bear *partial* ex post causal responsibility for event e_1. This is the case of the two independently successful research groups which each find a cure for cancer. Next, consider agent 1 taking action d. In this case, agent 1 is never pivotal for event e_1 for any of agent 2's possible actions. Hence, for both possible realized action profiles the ex post causal responsibility of agent 1 for event e_1 is zero,

[13] Note that, in principle, for every action profile, the evaluation of four different ex post causal responsibility values is possible - one for each player-event combination.

$r^{EP}_{1,e_1}(d,l) = r^{EP}_{1,e_1}(d,r) = 0$. Agent 1 is said to have *no* ex post causal responsibility for event e_1.

Substitutes environment

	l	r
u	e_1 $r^{EP}_{1,e_1}(u,l) = \frac{1}{2}$	e_1 $r^{EP}_{1,e_1}(u,r) = 1$
d	e_1 $r^{EP}_{1,e_1}(d,l) = 0$	e_2 $r^{EP}_{1,e_1}(d,r) = 0$

Complements environment

	l	r
u	e_1 $r^{EP}_{1,e_1}(u,l) = 1$	e_2 $r^{EP}_{1,e_1}(u,r) = \frac{1}{2}$
d	e_2 $r^{EP}_{1,e_1}(d,l) = 0$	e_2 $r^{EP}_{1,e_1}(d,r) = 0$

Table 3.2: Ex post causal responsibility in two simple environments.

Next, consider the *complements* environment. This environments highlights two additional features of ex post causal responsibility. First, it distinguishes the notion of ex post causal responsibility from notions of diffusion of responsibility that simply split responsibility among the involved agents (Latané and Darley, 1968). Consider the action profile (u,l). In this case, agent 1, but also agent 2 are pivotal for event e_1 and, thus, both possess full ex post causal responsibility for the event. Ex post causal responsibility is therefore not diffused, if agents act as complements, but it is diffused, if agents act as substitutes. Second, the environment also captures how ex post causal responsibility can be evaluated for hypothetical events. Consider the action profile (u,r). In this case, event e_2 is implemented. However, agent 1 nevertheless bears partial ex post causal responsibility for the *unrealized* event e_1, for which his action would have been pivotal, if agent 2 had played l instead of r. If agent 1 chooses action d, he bears no ex post causal responsibility for event e_1, similar to the *substitutes* environment.

3.2.2 Ex ante causal responsibility

While the notion ex post causal responsibility has many appealing features, it alone cannot capture all intricacies that we expect to play a role for the overall assessment of an agent's causal responsibility for an event. In particular, one challenge arises which makes it necessary to complement the notion of ex post causal responsibility with an *ex ante analysis*. Namely, two agents can both be ex post causally responsible to the same degree for an event even if, ex ante, their actions had vastly different probabilities to reach the same level of ex post causal responsibility. For example, compare two agents who each decide whether to shoot a gun or not. The only difference being that if they

shoot, the first agent has a 99 percent chance to kill an innocent bystander and the second has a 1 percent chance to kill an innocent bystander. Suppose both agent shoot and a bystander is killed in both cases. In this case, both agents would be attributed full ex post causal responsibility for the death of the bystander even if, ex ante, their actions had vastly different probabilities to reach that level of ex post causal responsibility. The ex ante causal responsibility component handles such situations.

Let $\mathbf{a}_{-0} = (a_i)_{i \in I \setminus \{0\}}$. An agent i's ex ante causal responsibility for an event x is defined as the expected level of ex post causal responsibility for that event, given the actions of all agents (excluding nature) and nature's exogenously given probability distribution, ω.[14]

Definition 3. *An agent i's degree of ex ante causal responsibility for event $x \in X$ is defined as*

$$r_{i,x}^{EA}(\mathbf{a}_{-0}, \omega) = E_{\omega}[r_{i,x}^{EP}(\mathbf{a})]. \tag{3.3}$$

The ex ante causal responsibility component thus captures the fact that, in the presence of moves of nature, the agents' chosen actions induce an explicit probability distribution over ex post causal responsibility levels. Agents whose actions induce a higher level of expected ex post causal responsibility bear a higher level of ex ante causal responsibility. If there are no moves of nature, or, in other words, if nature's action set is a singleton, ex ante and ex post causal responsibility coincide.

Example

Consider an individual decision problem with risk in which an agent 1's action space is the choice of an integer from 0 to 9, $a_1 \in \{0, ..., 9\}$. Simultaneously, a lottery also picks an integer from 0 to 9 with 10 percent probability each. This probability distribution is denoted by ω. Event e_1 is implemented if the sum of the chosen numbers is at least 10, otherwise event e_2 is implemented. Thus, the agent can implement event e_2 alone, by choosing 0, but to implement event e_1, both, agent and lottery are needed.

What is the agent's ex post and ex ante causal responsibility for event e_1? First, the agent bears full ex post causal responsibility for event e_1 whenever it is implemented. This is because the agent could always prohibit the implementation of event e_1 by choosing 0. If the agent chooses a positive integer, but the lottery chooses such that event e_2 is implemented, then the agent has a degree of ex post causal responsibility of 1/2 for the hypothetical event e_1. He bears no ex post causal responsibility for event e_1, if he chooses

[14] I restrict ex ante causal responsibility to exclusively depend on randomness that is generated due to moves of nature. The reasons for this restriction are elaborated on in the next section, when the notion is introduced in a game-theoretic framework.

0, as then he could not be pivotal for event e_1 for any of the lottery's possible picks.

$$r^{EP}_{1,e_1}(a_1, a_0) = \begin{cases} 1 & \text{if } a_1 > 0 \text{ and } a_1 + a_0 \geq 10 \\ \frac{1}{2} & \text{if } a_1 > 0 \text{ and } a_1 + a_0 < 10 \\ 0 & \text{if } a_1 = 0 \end{cases}$$

This example illustrates the shortcoming of a measure that is solely based on an ex post notion of causal responsibility. The agent bears full ex post causal responsibility for event e_1 when he picks 1 and the lottery 9 and when he picks 9 and the lottery 1, even though the ex ante probabilities of reaching such levels of ex post causal responsibility, given his action, were very different.

The notion of ex ante causal responsibility captures just that. Let F_0 be the cumulative distribution function of the lottery. The ex ante causal responsibility of the agent for event e_1 is

$$r^{EA}_{1,e_1}(a_1, \omega) = \begin{cases} (1 - F_0(10 - a_1)) \cdot 1 + F_0(10 - a_1) \cdot \frac{1}{2} = \frac{a_1}{10} \cdot 1 + (1 - \frac{a_1}{10}) \cdot \frac{1}{2} & \text{if } a_1 > 0 \\ 0 & \text{if } a_1 = 0. \end{cases}$$

Ex ante causal responsibility of the agent for event e_1 thus increases in the agent's stated integer value.

3.2.3 Overall causal responsibility

The two notions of ex ante and ex post causal responsibility are now combined as a convex combination to yield a function of overall causal responsibility.

Definition 4. *An agent i's degree of overall causal responsibility for event $x \in X$ is defined as*

$$r_{i,x}(\boldsymbol{a}, \omega) = \alpha \cdot r^{EA}_{i,x}(\mathbf{a}_{-0}, \omega) + (1 - \alpha) \cdot r^{EP}_{i,x}(\mathbf{a}) \tag{3.4}$$

with $\alpha \in [0, 1]$.

The parameter α is an individual-specific parameter of the agent who evaluates the causal responsibility of agent i for event x. An agent with $\alpha = 0$ places value on ex ante causal responsibility only, and might therefore be understood as an agent who is only driven by deontological motives. An agent with $\alpha = 1$, on the other hand, only considers ex post causal responsibility, and might be understood as an agent who is only driven by

consequentialist motives. When $\alpha \in (0,1)$ a combination of ex ante and ex post causal responsibility is used. Oftentimes, this seems to be the most natural case. In the example from above, $\alpha \in (0,1)$ would mean that for the evaluation of causal responsibility for killing another person by shooting a gun, it matters, both, whether a person was actually killed (ex post causal responsibility) and with which probability a shot would result in the death (ex ante causal responsibility). An appropriate experimental design is able to test whether both factors play a role and allows the structural estimation of the parameter α.

An important feature of the notion is that it can potentially be included in different preference frameworks and different strategic settings. Therefore, it should be understood and used as a portable extension of existing models (PEEM) (Rabin, 2013). For example, if one is interested in studying settings of distributive justice, then including the notion of causal responsibility into a preference framework of inequity aversion can be used to study how agents react differently to inequality depending on how responsible they were for it. On the other hand, a principle who wants to reward agents for the successful implementation of a project can use causal responsibility to allocate a bonus in proportion to the agents' causal responsibility for the project. In the following, I will incorporate the notion in a framework of *responsibility preferences* in which an agent has a preference to reward and punish other agents for being causally responsible for events that he deems bad or good.

3.3 Model

After having defined the notion of causal responsibility, it can now be used to study how responsibility perceptions influence behavior. To this end, I incorporate evaluations of causal responsibility into the utility function. Specifically, I assume that there exists an agent with *responsibility preferences*. He observes the behavior of other agents and has a taste, in addition to his taste for monetary payoff, to reward or punish those other agents for the implementation of what he judges as good or bad events, but only to the extent those agents are causally responsible for them.

I analyze the behavioral implications of *responsibility preferences* in a two-stage game. Stage-1, the *collective action stage*, is a simultaneous-move game in which a group of agents, the stage-1 agents and, potentially, nature, simultaneously take actions. These actions generate a stage-1 payoff for the stage-1 agents and, together, also result in an *event*. The stage-1 agents have preferences over monetary payoff only. Another agent, who is not involved in its implementation, cares about which event is implemented. For example, the stage-1 agents could be thought of as a group of workers, each of whom decides whether to work or to shirk. The workers only care to maximize their utility.

The boss, on the other hand, cares about whether the group finishes a team project successfully, or not (event).

In stage-2, the *responsibility attribution stage*, the affected agent judges the possible events in stage 1 and the stage-1 agents' causal responsibility for them. He is assumed to possess *responsibility preferences* and will therefore assign punishment and reward to the stage-1 agents in relation to their responsibility for what he judges as good or bad stage-1 events. For example, the boss will want to reward and punish those workers that were causally responsible for the success or failure of the team project.

3.3.1 Setup

Formally, the game is set up as follows:

Stage 1 - *Collective action stage*
In stage 1, each of a finite set of agents $I = \{0, 1, ..., n\}$ simultaneously takes an actions $a_i \in A_i$ where A_i denotes agent i's finite set of feasible actions. If present, nature is denoted as agent 0. The set $A = \prod_{i \in I} A_i$ is the set of all feasible action profiles in stage 1 and $\mathbf{a} = (a_0, ..., a_n)$ is an element of that set. A behavioral strategy of agent i in stage 1, denoted by σ_i, is a probability distribution over the agent's action set A_i and nature's "strategy" ω is an exogenously given probability distribution with full support over A_0, which is common knowledge. The set of feasible strategies of agent $i \in I \setminus \{0\}$ is denoted by $\Sigma_{i \in I \setminus \{0\}} = \Delta A_{i \in I \setminus \{0\}}$ and the set of feasible strategy profiles $\Sigma = \prod_{i \in I \setminus \{0\}} \Sigma_i$.

The function $\pi_i^I : A \to \mathbb{R}$ links the stage-1 action profile to the monetary stage-1 payoff of each agent $i \in I \setminus \{0\}$. Additionally, there exists an agent $K \notin I$, who is inactive in stage 1, but will become active in stage 2. The function $f : A \to X$ links the action profile to the event x. Agent K cares about the event, which is captured by his stage-1 payoff function $\pi_K^I : X \to R$.[15] At the end of stage 1, the realized action profile $\mathbf{a}$ and the resulting payoffs are known to all agents. I denote the history of play after stage 1 by $h \in H$.

Stage 2 - *Responsibility attribution stage*
In stage 2, only agent K takes an action, which determines the stage-2 payoffs. Specifically, after each history h, agent K makes an allocation decision $p_i(h) \in P_i$ for each stage-1 agent $i \in I \setminus \{0\}$, where P_i is the set of feasible, history-independent allocations for agent i. Agent K's action space at history h is thus $P = \prod_{i \in I \setminus \{0\}} P_i$ and the n-dimensional vector

[15] Note that the restriction to a single stage-2 agent is for notational simplicity only and that π_K^I is a payoff expressed in monetary units which does not necessarily equal his monetary payoff. For example, agent K could also be a third party that cares about the payoff distribution between the stage-1 agents and a second party.

$\mathbf{p}(h) = [p_1(h), ..., p_n(h)]$ denotes an element of the set. A behavioral strategy for agent K is a function σ_K that associates with every history $h \in H$ a probability distribution $\sigma_K(h)$ over P. The set of feasible strategies is $\Sigma_K = \Delta P$.

The stage-2 monetary payoffs after history h are denoted by the functions $\pi_i^{II} : A \times P \to \mathbb{R}$ for stage-1 agents $i \in I \setminus \{0\}$ and $\pi_K^{II} : A \times P \to \mathbb{R}$ for agent K. Let $z \in Z \subset H$ denote the terminal history and Z the set of feasible terminal histories. Stage-1 agents are assumed to possess monotonic preferences over monetary payoffs only. Their stage utility is thus $u_i^I(\mathbf{a}) = \pi_i^I(\mathbf{a})$ and $u_i^{II}(p_i(h)) = \pi_i^{II}(p_i(h))$. The function $U_i : Z \to \mathbb{R}$ denotes the utility of agent i from the game as a whole, which is simply defined as the sum of his utilities in the two stages (no discounting). The stage-1 agents choose strategies to maximize their expected utility from the game as a whole, rationally anticipating the behavior of agent K in stage 2.

3.3.2 Responsibility preferences

Agent K, on the other hand, is assumed to possess *responsibility preferences*. He observes the actions of the agents in stage 1, the implemented event, and the resulting stage-1 payoffs. He has a preference, in addition to his preference for monetary payoff, to reward or punish stage-1 agents for the implementation of what he judges as good or bad events in stage-1, but only to the extent those agents are causally responsible for them. In the following, I introduce the components of a utility function that represents such preferences.

I assume that agent K judges the possible stage-1 events $x \in X$ according to a *judgment function* $j : X \to \mathbb{R}$. Agent K can judge an event x as *good* ($j(x) > 0$), *bad* ($j(x) < 0$), or *neutral* ($j(x) = 0$). Specifically, I assume that the judgment depends on the payoff that an event generates for agent K in stage 1, $\pi_K^I(x)$, relative to a reference payoff, $\bar{\pi}_K^I(X)$, which depends on the set of feasible payoffs.[16]

Definition 5. *Agent K's judgment of event x is given by the judgment function* $j : X \to \mathbb{R}$ *which is defined as*

$$j(x) = \pi_K^I(x) - \bar{\pi}_K^I(X) \tag{3.5}$$

with $\bar{\pi}_K^I(X) \in [\min_{x \in X} \pi_K^I(x), \max_{x \in X} \pi_K^I(x)]$.

Formulating the judgment function in relative terms has two appealing features. First, when only two events are possible, they are judged neutrally only in case of indifference, when both generate the same payoff. This captures the fact that we don't reward two

[16] For example, a reference payoff equal to the average between the best and the worst possible payoff, $\bar{\pi}_K^I(X) = 0.5 \cdot [\min_{x \in X} \pi_K^I(x) + \max_{x \in X} \pi_K^I(x)]$ would be similar to intention-based social preference models (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004).

equally "good" or punish two equally "bad" events, when nothing else was possible. Second, it provides an intuitive scale: The more an event's payoff deviates from the reference payoff, the better or worse it is judged to be.[17]

Agent K then evaluates the causal responsibility, $r_{i,x}(\mathbf{a}, \omega)$, of each of the stage-1 agents for each possible event, as introduced in section 2. Note that, in this strategic setting, ex ante causal responsibility could, in principle, depend on two types of uncertainty: objective uncertainty, stemming from moves of nature, and subjective uncertainty, stemming from the stage-1 agents' behavioral strategies. Therefore, one might argue that agent K's beliefs about the behavioral strategies of the stage-1 agents should be accounted for in the ex ante causal responsibility component instead of the realized stage-1 action profile. I follow Dufwenberg and Kirchsteiger (2004) and Battigalli and Dufwenberg (2009)'s formulation of dynamic psychological games and assume that agents play pure strategies only and that behavioral strategies represent the frequencies with which pure strategies are chosen in a population. Therefore, opposed to moves of nature, behavioral strategies do not induce actual randomization in stage-1 and should not be accounted for in the ex ante causal responsibility component (for similar arguments, see Sebald (2010) and Blanco, Çelen, and Schotter (2013)). Hence, after observing history h, agent K implicitly updates his beliefs to concord with the actions that lead to history h. Since, in stage-2, only updated beliefs enter the ex ante causal responsibility function, I abstain from modeling beliefs and the updating process explicitly and directly replace them with the observed actions.

Given how agent K judges the stage-1 events and evaluates the agents' causal responsibility for them, agent K's *overall judgment of the behavior of agent* i, given i's action, comprises the sum over the judgments of all possible events, weighted by agent i's causal responsibility for them, $\sum_{x \in X} r_{i,x}(\mathbf{a}, \omega) \cdot j(x)$. Similar to the judgment of events, behavior can be judged as *praiseworthy* ($\sum_{x \in X} r_{i,x}(\mathbf{a}, \omega) \cdot j(x) > 0$), *blameworthy* ($\sum_{x \in X} r_{i,x}(\mathbf{a}, \omega) \cdot j(x) < 0$), or *neutral* ($\sum_{x \in X} r_{i,x}(\mathbf{a}, \omega) \cdot j(x) = 0$). Note that the overall judgment of agent i's behavior is sensitive to the reference payoff, $\bar{\pi}_K^I(X)$. Three points, in particular, are worth highlighting: First, if the reference payoff is equal to the payoff of one of the possible events, e.g., the event that generates the highest or lowest stage-1 payoff for agent K, then that event is judged as *neutral*. Any level of causal responsibility for a neutral event does not change the overall judgment of agent i's behavior, as $r_{i,x}(\mathbf{a}, \omega) \cdot 0 = 0$. Second, an interior reference payoff, $\bar{\pi}_K^I(X) \in (\min_{x \in X} \pi_K^I(x), \max_{x \in X} \pi_K^I(x))$, leads to

[17] The reference payoff, $\bar{\pi}_K^I(X)$, could be interpreted as an individual- and context-specific parameter of agent K which can be identified with appropriate experimental methods. For example, the share given to a recipient in a dictator game at which a third party switches from punishing to rewarding the dictator can be interpreted as the neutral event that determines the reference payoff of the third party (Nikiforakis and Mitchell, 2013). Any share above is interpreted as a good, and any share below is interpreted as a bad event for that third party. This would capture the fact that individuals are heterogeneous in their evaluation of events. Some have high reference points and judge any deviation as bad, and others have low reference payoffs and judge any deviation as good.

some events being judged as bad ($j(x) < 0$) and others as good ($j(x) > 0$). The additive nature of the overall judgment of behavior induces the important feature that causal responsibility for a bad event and simultaneous causal responsibility for a good event can (partly) cancel each other out. For example, let $j(e_1) > 0 > j(e_2) > j(e_3)$. If agent i is fully causally responsible for event e_3, then agent K's overall judgment of i's behavior will increase if i is also partially causally responsible for event e_1, but decrease if he is also partially causally responsible for event e_2.[18] However, the additive nature also implies that the relative pattern of overall judgment can be sensitive to the reference payoff. For example, take an agent who is fully causally responsible for event e_3 and, in addition, partially causally responsible for event e_2. Whether the partial causal responsibility for event e_2 increases or decreases the overall judgment of his behavior by agent K depends on whether the reference payoff of K is lower or higher than the payoff of event e_2, which determines whether $j(e_2) > 0$ or $j(e_2) < 0$. This potential sensitivity of the overall judgment to the reference payoff makes it imperative to discuss any assumption made on the reference payoff and its potential impact on the analysis. It does not mean, however, that there always exists a reference payoff that explains any findings and, thus, that the theory is not falsifiable. In all examples discussed and all experiments analyzed in this paper, the reference payoff, as long as it is interior, has no influence on the comparative static predictions of the theory.

Given agent K's overall judgment of agent i's behavior after history h, he chooses an allocation, $p_i(h)$, for that agent accordingly. Agent K is said to *punish* agent i, if he reduces his stage-2 payoff ($p_i(h) < 0$), and he is said to *reward* agent i, if he increases his stage-2 payoff ($p_i(h) > 0$).[19] We are now ready to define the stage-2 utility function of agent K with responsibility preferences.

Definition 6. *The stage-2 utility of agent K is a function $u_K^{II} : A \times P \times \{\omega\} \to \mathbb{R}$ that is defined as*

$$u_K^{II}(\mathbf{p}(h), h, \omega) = \pi_K^{II}(\mathbf{p}(h)) + \rho \sum_{i \in I \setminus \{0\}} \left[\sum_{x \in X} r_{i,x}(\mathbf{a}, \omega) \cdot j(x) \right] \cdot p_i(h) \tag{3.6}$$

The parameter $\rho \geq 0$ captures how much agent K cares about punishing or rewarding the behavior of the stage-1 agents compared to monetary payoff. To maximize his utility, agent K will match the signs of the overall judgment of behavior of agent i and the allocation to agent i. Therefore, an overall blameworthy (praiseworthy) behavior of agent i is matched with punishment (reward) of agent i.

[18] See Section 3.4.2 for experimental evidence on this additive nature.

[19] The implicit reference allocation is therefore zero, the allocation that neither in- nor decreases agent i's payoff. This is the natural allocation reference when thinking about punishment and reward.

The overall utility of agent K is a function $U_K : Z \times \{\omega\} \to \mathbb{R}$ and thus depends on the terminal history and nature's "strategy" in stage 1. It is simply defined as the sum of the payoffs from stage 1 and 2 (no discounting). The game is thus specified as $\Gamma = (I \cup K, A, P, \omega, (U_i)_{i \in I \setminus \{0\}}, U_K)$. Since this is a standard multi-stage game with moves of nature, all standard equilibrium concepts apply and a subgame perfect equilibrium is guaranteed to exist.

3.3.3 A simple example

In this section, I use a simple numerical example to demonstrate the implications of *responsibility preferences* and the analysis of subgame perfect equilibria. Throughout the analysis, for ease of exposition, I will focus on pure strategy equilibria only. In the example, two politicians, $I = \{A, B\}$, vote in congress to enact a bill into law by voting "yay" or "nay", $A_i = \{y, n\} \ \forall \ i \in I$. The bill is passed, S, if at least one of them votes for it, and it fails, F, otherwise. Thus, the set of events is $X = \{S, F\}$. If the bill passes, both politicians' remuneration is raised, but at the expense of the electorate, which is represented by agent K. Table 3.3 provides the stage-1 payoffs of the game. Note the game resembles the environments discussed in section 2: politicians act as substitutes in implementing S and as complements in implementing F. At the end of stage 1, the actions, the implemented event, and the resulting payoffs are known to all agents, matching a scenario in which politicians' votes can be observed (e.g. votes of the US congress).

	y	n
y	$(5, 5, 1)$	$(5, 5, 1)$
n	$(5, 5, 1)$	$(3, 3, 5)$

Table 3.3: Stage-1 payoffs. The left, middle, and right number are agent A, B, and K's payoffs.

In stage 2, after history h, agent K can choose an allocation $p_i(h) \in [-10, 10]$ for each of the two politicians $i \in \{A, B\}$, which might be interpreted as the electorate's support in the next election. The allocation choice for agent i induces his stage-2 payoff $\pi_i^{II}(p_i(h)) = p_i(h)$. The vector of allocation decisions, $\mathbf{p}(h) = [p_A(h), p_B(h)]$ induces the stage-2 payoff of agent K after history h, $\pi_K^{II}(\mathbf{p}(h)) = -\sum_{i \in I} \frac{p_i(h)^2}{2}$. The convex cost function ensures an interior solution and can be interpreted as the increasing marginal cost of opposing or supporting a given politician.

I assume that the reference payoff for the judgment of the law for agent K is $\bar{\pi}_K^I(X) = 3$, the midpoint between the payoffs he receives if the law passes or fails. Note that, with only two possible events, the prediction for the qualitative allocation pattern is identical for any interior reference payoff $\bar{\pi}_K^I(X) \in (\pi_K^I(S), \pi_K^I(F))$. Agent K thus judges the failure of the law as a good stage-1 event $j(F) = 5 - 3 = 2 > 0$ and the passage as a bad stage-1 event $j(S) = 1 - 3 = -2 < 0$. Because there are no moves of nature involved, agent K will judge the causal responsibility of the politicians only according to their chosen actions, similar to Table 3.2. The first-order condition of utility maximization of agent K in stage 2 yields the best-response function of agent K after history h for politician A:

$$p_A^*(h) = \begin{cases} \rho \, \frac{1}{d_{A,S}(y,y)} \, j(S) = \rho \cdot \frac{1}{2} \cdot (-2) = -\rho & \text{if} \quad h = (y,y) \\ \rho \, \frac{1}{d_{A,S}(y,n)} \, j(S) = \rho \cdot 1 \cdot (-2) = -2\,\rho & \text{if} \quad h = (y,n) \\ \rho \, \frac{1}{d_{A,F}(n,y)} \, j(F) = \rho \cdot \frac{1}{2} \cdot 2 = \rho & \text{if} \quad h = (n,y) \\ \rho \, \frac{1}{d_{A,F}(n,n)} \, j(F) = \rho \cdot 1 \cdot 2 = 2\,\rho & \text{if} \quad h = (n,n) \end{cases}$$

The allocation for politician B is symmetric. Thus, *responsibility preferences* yield a clear pattern of allocation decisions. For example, if $\rho = 0.5$, then $p_A^*(y,n) = -1$, $p_A^*(y,y) = -0.5$, $p_A^*(n,y) = 0.5$, and $p_A^*(n,n) = 1$. In a subgame perfect Nash equilibrium, the politicians rationally anticipate the behavior of the electorate in stage 2 and maximize their payoffs from the game as a whole, which are presented in Table 3.4.

	y	n
y	$5-\rho$ $5-\rho$	$5-2\,\rho$ $5+\rho$
n	$5+\rho$ $5-2\,\rho$	$3+2\,\rho$ $3+2\,\rho$

Table 3.4: Overall payoffs for agent A and B.

It is easy to see that playing (y,y) in stage 1 cannot be part of an equilibrium as, even though (y,y) diffuses causal responsibility among the two politicians, both have an incentive to deviate and vote n in order to avoid being opposed, while still enjoying the high payoff in stage 1. If $\rho \leq \frac{1}{2}$, there exist two subgame perfect equilibria in which exactly one politician votes "yay", one politician votes "nay" ((y,n) and (n,y)), the law passes and the electorate allocates support and opposition according to $p_i^*(h)$. On the other hand, if $\rho > \frac{1}{2}$, then there exists a unique subgame perfect equilibrium in which both politicians vote "nay" (n,n), the law fails to be passed, and the electorate allocates support and opposition according to $p_i^*(h)$.

Hence, if the electorate does not care enough about allocating punishment and reward for the passage or failure of the law, it will be implemented with just enough votes as necessary. If, however, the electorate cares enough, the law will not be implemented as the opposition a politician incites by implementing the law is enough to deter him from voting for it. Generalized to more than two politicians, the results remain the same. In equilibrium, such a law is either implemented with the smallest majority necessary, or, if it is not implemented, everyone votes against it. This tracks nicely the anecdotal evidence which suggests that political parties often have problems motivating their own members to vote for a particular issues because the party members fear they will be held responsible in their constituency. In fact, there even exists an official position, the "party whip", whose task is to find majorities and ensure party discipline in legislative voting. One example for such equilibria in the real world is Germany's politicians' stance on nuclear energy. For some time, there was a small majority in parliament that was in favor of nuclear energy ((y, n) equilibrium) . After Fukushima, public awareness increased (ρ increased), and the parliament voted for Germany's exit from nuclear energy. Since then, no politician openly argues in favor of nuclear energy ((n, n) equilibrium).

3.3.4 Comparison with alternative theories

At this point, it is instructive to discuss the predictions of other preference frameworks. Certainly, a new theory is only of interest if it makes distinctively different predictions than other, competing theories, in situations in which the predictions of the established theories seem to go wrong. In the following, I will discuss the predictions of other theories for the simple example introduced above.

Standard preferences

The workhorse model of standard preferences assumes that agents act in order to maximize their monetary payoff. If this is the case, agent K would never deviate from the allocation $p_i(h) = 0$ after any history h as doing so is costly. In stage 1, the two agents rationally anticipate that they will neither get punished nor rewarded in stage 2. Therefore, (n, n) is never played in equilibrium in stage 1 and the law always passes. In stage 1, the action profiles (y, y), (y, n), and (n, y) can all be part of an SPNE . Hence, in comparison to *responsibility preferences*, standard preferences predict neither punishment nor reward, the law is always passed and also (y, y) is part of an equilibrium.

Outcome-based social preference model

Models of outcome-based social preferences allow concerns about the payoff distribution to enter the utility function. For example, the two prominent theories by Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) assume that agents have an aversion against

final payoff inequalities. If agent K possesses outcome-based social preferences in the formulation of Fehr and Schmidt, he chooses $p_i(h)$ to maximize his utility from the game which is given by

$$U_K(h, \mathbf{p}(h)) = \pi_K^I(h) - \sum_{i \in I} \frac{p_i(h)^2}{2} - \alpha \sum_{i \in I} \max[(\pi_i^I(h) + p_i(h)) - (\pi_K^I(h) - \sum_{i \in I} \frac{p_i(h)^2}{2}), 0]$$
$$- \beta \sum_{i \in I} \max[(\pi_K^I(h) - \sum_{i \in I} \frac{p_i(h)^2}{2}) - (\pi_i^I(h) + p_i(h)), 0]$$

The parameters $\alpha > \beta \geq 0$ measure how much agent K cares about advantageous (β) and disadvantageous (α) inequality. Agent K uses punishment and reward to reduce payoff inequalities between him and the stage-1 agents. Therefore, he will punish both politicians if the law is passed and reward both politicians if the law fails, but never more as to achieve payoff equality. In case of an interior solution (agent K would not punishment or reward more than to achieve equality of payoffs), agent K allocates after history h according to

$$p_i^*(h) = \begin{cases} \frac{-\alpha}{(1+\alpha)} & \text{if} \quad h \in \{(y,y),(y,n),(n,y)\} \\ \frac{\beta}{(1-\beta)} & \text{if} \quad h = (n,n) \end{cases}$$

Importantly, punishment and reward only depend on the payoffs that were implemented in stage 1 and are independent of the stage-1 actions that led to those payoff. Therefore, if the law is passed, agent K will punish agent A and B equally, independent of whether only one or both of them voted "yay". Similarly, when the law fails, reward will be equal for both stage-1 agents. Note that (y, y) is always sustained as part of a subgame perfect Nash equilibrium. This is because any deviation of one of the stage-1 agents would not alter his payoff. The law would still be passed and punishment in stage-2 would still be the same. Whether (n, y) and (y, n), or (n, n) are part of a SPNE depends on the specific parameters. Thus, the two important differences to the *responsibility preference* framework are that a) the two stage-1 agents are always punished and rewarded equally, only depending on which event is implemented and b) (y, y) is always sustained as a SPNE.

Intention-based social preferences

Another class of social preferences focuses on intention-driven reciprocity. The key assumption of these models is that kind behavior is reciprocated with kind behavior and that unkind behavior is reciprocated with unkind behavior (Rabin, 1993). Given the sequential nature of the example, Dufwenberg and Kirchsteiger (2004) is applicable.

As before, I assume that stage-1 agents act only to maximize their material payoff. Furthermore, I assume that agent K evaluates the kindness of agent A and B only based on

his stage-1 payoff, and that therefore any alterations of his payoff due to the costs he bears for punishment and reward in stage 2 do not influence his evaluation of kindness. Then, agent K perceives agent A's action as kind (unkind) if K believes that A believes that his action grants K a higher (lower) payoff than certain reference payoff. The reference payoff is assumed to be the average of the lowest and the highest efficient payoff that A could give K, according to his beliefs about agent K and agent B's strategy.

Agent K will punish (reward) a stage-1 agent, if that agent granted him a payoff that is lower (higher) than the reference payoff. The allocation predictions for agent A are as follows: If agent B votes "yay", then the bill is passed independently of agent A's vote. Hence, in this case, agent A cannot influence agent K's payoff and no matter whether he votes "yay" or "nay" his action is judged as neither kind nor unkind and will therefore neither incite punishment nor reward in stage 2. On the other hand, if agent B votes "nay", then agent B can influence whether the bill is passed or not and therefore the payoff of agent K in stage 1. As, in this case, voting "nay" would grant agent K a higher payoff than voting "yay", voting "yay" is interpreted as an unkind action and voting "nay" is interpreted as a kind action. Therefore, in stage 2, agent K will reward agent A for voting "nay" and punish agent A for voting "yay". The allocation predictions for agent B are symmetric.

Anticipating the best-response of agent K in stage 2, in stage 1, playing (y, y) is always sustained as part of a SPNE because deviating neither changes stage-1 payoffs nor agent K's perception of one's kindness. Additionally, depending on how much agent K cares about reciprocation ($\rho > 0$), either (n, n) or (n, y) and (y, n) are possible as part of an equilibrium. Thus, the main difference to *responsibility preferences* is that intention-based social preferences make no differentiated allocation predictions once an agent's action is not pivotal and thus is not fully causally responsible. Therefore, intention-based models fail to explain many situations that are of interest in group decision making.

Model / History	(y, y)	(y, n)	(n, y)	(n, n)
Standard	0	0	0	0
Causal responsibility	$-\rho$	$-2\,\rho$	$+\rho$	$+2\,\rho$
Inequity aversion	$-\frac{\alpha}{(1+\alpha)}$	$-\frac{\alpha}{(1+\alpha)}$	$-\frac{\alpha}{(1+\alpha)}$	$+\frac{\beta}{(1-\beta)}$
Reciprocity	0	$-\rho$	0	$+\rho$

Table 3.5: Agent K's optimal allocation to agent A under different preference frameworks.

Table 3.5 summarizes the predictions of the different preference models for the allocation decision of agent K in stage 2 after all possible histories. It shows that the responsibility-based preference model makes distinctive predictions compared to the other theories. Interestingly, responsibility-based preference-theory is the only theory that does not support (y, y) as part of a subgame-perfect Nash equilibrium, which, intuitively, makes sense.

By deviating from (y, y) an agent's stage-1 payoff is kept constant, but he can hope to get a higher payoff in stage-2. In the responsibility framework, such a higher payoff is generated, because, by deviating in stage-1, an agent can shed himself from being causally responsible for the event that agent K dislikes.

3.3.5 The effects of causal responsibility attribution in two common environments

While the last section served to demonstrate the workings of causal responsibility-based preferences and the differences to other theories in a simple numerical example, this section goes a step further and analyzes how causal responsibility attribution affects equilibrium outcomes in more generalized environments. Specifically, I analyze two environments in which everything is identical but the payoff structure of the stage-1 agents in stage 1.

In the first environment, the *private gains* environment, stage-1 agents can choose between two actions. One leads to a private benefit, but, if enough stage-1 agents take it, an event occurs that is detrimental to agent K. Thus, this is an environment with private gains and a potential negative externality. For example, one can think of workers who derive private gain from shirking, firms that can increase their profit by choosing a dirty technology, or airplane passengers that gain from taking the plane. If enough workers shirk, the project fails, if enough firms use the dirty technology, the environment is destroyed, and if enough passengers book a flight, the plane flies and creates CO2 emmisions. Each of these events is judged as bad by the stage-2 agent, who can be thought of as a boss who dislikes the failure of the team-project, a consumer who dislikes destruction of the environment, or society, which dislikes climate change.

In the second environment, the *collective gains* environment, stage-1 agents again choose between two actions. The only difference is that, in this environment, a certain number of stage-1 agents has to take an action in order to create a benefit for *all* stage-1 agents. Again, if enough of them take the action, an event is triggered that is detrimental to agent K. For example, one could think of politicians who can only implement a reform that benefits themselves at the cost of the taxpayer if enough of them vote for it, or of firms which can only form a cartel if enough of them participate, but, if the cartel is established, the higher price benefits all firms. Agent K can be thought of as the electorate, that doesn't like if politicians enrich themselves, or of consumers that have to pay higher prices.[20]

[20]The *collective gains* environment is thus a generalization of the simple numerical example in the last section.

The formal setup is as follows: In both environments, the set of stage-1 agents, I, is indexed by $i \in \{1, ..., n\}$, nature is not present, and an agent $K \notin I$ acts only in stage 2. The stage-1 agents' identical action sets consists of two actions, $A_i = \{\underline{a}, \overline{a}\}\ \forall\ i \in I$. The stage-1 actions induce one of two possible events, $X = \{\underline{e}, \overline{e}\}$. The function f, which governs which event is implemented, is defined as

$$f(\mathbf{a}) = \begin{cases} \underline{e} & \text{if} \quad \sum_{i \in I} \mathbb{1}(a_i = \underline{a}) \geq t \\ \overline{e} & \text{if} \quad \sum_{i \in I} \mathbb{1}(a_i = \underline{a}) < t \end{cases} \tag{3.7}$$

where $n > t > 0$ and $\mathbb{1}(a_i = \underline{a}) = 1$ if $a_i = \underline{a}$ and zero otherwise. Thus, when at least t stage-1 agents choose action $\underline{a}$, event $\underline{e}$ is implemented.

In stage 1, agent K gets a strictly higher payoff, if event $\overline{e}$ is implemented, $\pi_K^I(\overline{e}) > \pi_K^I(\underline{e})$. The payoffs of the stage-1 agents, however, depend on the specific environment. In the *private gains* environment, their stage-1 payoff only depends on their individual action and their payoff is higher, if they take action $\underline{a}$, $\pi_i^I(\underline{a}) > \pi_i^I(\overline{a})\ \forall\ i \in I$. Thus, the agents privately benefit from taking $\underline{a}$, but if at least t of them do so, event $\underline{e}$, which hurts agent K, is implemented. In the *collective gains* environment, on the other hand, all stage-1 agents get a higher payoff only if enough of them take action $\underline{a}$ to implement event $\underline{e}$. Thus, $\pi_i^I(\underline{e}) > \pi_i^I(\overline{e})\ \forall\ i \in I$.

At the end of stage 1, agent K observes the actions that were taken and the event $f(\mathbf{a})$ that was implemented. In stage 2, after history h, agent K chooses an allocation for each stage-1 agent, $p_i(h) \in [\underline{p}, \overline{p}]\ \forall\ i \in I$, with $\overline{p} > 0 > \underline{p}$. The allocation choice for agent i induces agent i's stage-2 payoff $\pi_i^{II}(p_i(h)) = p_i(h)$, which simply equals the allocation. Agent K's stage-2 payoff is determined by his allocation decisions. Specifically, he is assumed to face convex allocation costs, such that his stage-2 payoff equals $\pi_K^{II}(\mathbf{p}(h)) = -c \sum_{i \in I} \frac{p_i(h)^2}{2}$, where $c > 0$ is a cost parameter.

We can solve both games for the subgame perfect Nash equilibria by applying backward induction. As event $\overline{e}$ grants him a higher stage-1 payoff than event $\underline{e}$, agent K judges event $\overline{e}$ as better than event $\underline{e}$, $j(\overline{e}) > j(\underline{e})$. This holds independent of the reference point. Importantly, a stage-1 agent who chooses $\underline{a}$ has positive causal responsibility for event $\underline{e}$ only and an agent who chooses $\overline{a}$ has positive causal responsibility for event $\overline{e}$ only. Under the assumption that agent K possesses responsibility preferences represented by a utility function of the form of equation 3.6, solving for the optimal allocation policy after history h yields

$$p_i^*(h) = \begin{cases} \frac{\rho}{c} \frac{1}{d_{i,\underline{e}}(\underline{a}, a_{-i})}\ j(\underline{e}) & \text{if} \quad a_i = \underline{a} \\ \frac{\rho}{c} \frac{1}{d_{i,\overline{e}}(\overline{a}, a_{-i})}\ j(\overline{e}) & \text{if} \quad a_i = \overline{a} \end{cases} \qquad \forall\ i \in I \tag{3.8}$$

Note that this best-response function is independent of the specific environment.[21] It can therefore always be used for the analysis of causal responsibility attribution in environments with binary events and binary actions. In stage 1, the stage-1 agents rationally anticipate the optimal allocation policy and choose their strategies in order to maximize their expected payoff from the whole game (no discounting). In the following, I discuss the different equilibrium predictions in the *private gains* and the *collective gains* environment.

Proposition 1 (*Private gains* environment)**.**
(1) Suppose $\rho < c \cdot \frac{\pi_i^I(\underline{a}) - \pi_i^I(\overline{a})}{j(\overline{e}) - j(\underline{e})}$*. Then there exists a unique subgame perfect Nash equilibrium in which all stage-1 agents choose action* $\underline{a}$*, event* $\underline{e}$ *is implemented, and agent* K *allocates according to equation (3.8).*

(2) Suppose $\rho \geq c \cdot \frac{\pi_i^I(\underline{a}) - \pi_i^I(\overline{a})}{j(\overline{e}) - j(\underline{e})}$*. Then there exists an integer* $m \geq 1$ *for which it holds that:*[22]

i) If $t - m > 0$*, then there exists a SPNE in which* $t - m$ *stage-1 agents choose action* $\underline{a}$*,* $n - (t - m)$ *choose action* $\overline{a}$*, event* $\overline{e}$ *is implemented, and agent* K *allocates according to equation (3.8).*

ii) If $t - m \leq 0$*, then there exists a SPNE in which all stage-1 agents choose action* $\overline{a}$*, event* $\overline{e}$ *is implemented, and agent* K *allocates according to equation (3.8).*

iii) If $t + m - 2 < n$*, then there exists a SPNE in which all stage-1 agents choose action* $\underline{a}$*, event* $\underline{e}$ *is implemented, and agent* K *allocates according to equation (3.8).*

Proof. See Appendix 3.A.1.

The intuition behind the first part of the proposition is that if agent K does not care enough about the stage-1 agents behavior to deter them from taking $\underline{a}$ when doing so makes them fully causally responsible for event $\underline{e}$, then they will not be deterred for any level of causal responsibility for event $\underline{e}$ and, in equilibrium, all stage-1 agents take $\underline{a}$. In this case, the behavior of stage-1 agents in equilibrium coincides with that of the SPNE with standard preferences for which agent K would never engage in costly allocation decisions and all stage-1 agents take action $\underline{a}$.

Second, when agent K cares enough to deter taking $\underline{a}$ when it leads to full causal responsibility for $\underline{e}$, multiple equilibria are possible. The farther away from t is the number of

[21] Throughout this section I assume that the interval of possible allocations for i, $[\underline{p}, \overline{p}]$, is large enough such that equation 3.8 yields an interior solution.

[22] The threshold number m depends on the cost of punishment, c, the preference parameter, ρ, the difference in stage-1 payoffs, $\pi_i^I(\underline{a}) - \pi_i^I(\overline{a})$, and the difference in judgment, $j(\overline{e}) - j(\underline{e})$. For notational simplicity, I omit these variables when writing m.

stage-1 agents that take $\underline{a}$ (in both directions), the smaller is their causal responsibility for event $\underline{e}$ and, hence, the smaller is the relative gain in stage 2 that could be generated by playing $\overline{a}$ instead of $\underline{a}$. At the same time, the difference in stage-1 payoffs, $\pi_i^I(\underline{a}) - \pi_i^I(\overline{a})$, is fixed. Therefore, there exists a threshold number of moves away from t, denoted by m, for which it holds that, if $(t-m)$ agents take $\underline{a}$ then an agents who takes $\overline{a}$ would not want to deviate, but if $(t-m-1)$ agents take $\underline{a}$, he would want to deviate and also take $\underline{a}$. If exactly $t-m$ agents take $\underline{a}$, no stage-1 agent has an incentive to deviate, which constitutes an equilibrium. Similarly, if $t+m-2$ stage-1 agents take $\underline{a}$, then any agent who takes $\overline{a}$ has no incentive to deviate, but if $t+m-1$ agents take $\underline{a}$, causal responsibility is diffused enough such that any agent who takes $\overline{a}$ has an incentive to deviate and take $\underline{a}$. Therefore, if $t+m-2 < n$, there exists an equilibrium in which all stage-1 agents take $\underline{a}$.[23]

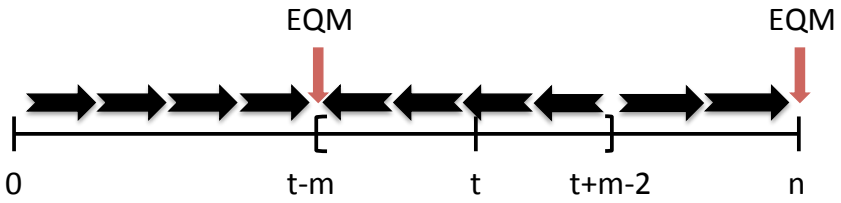


Figure 3.1: *Private gains* environment. The line shows the number of stage-1 agents that choose $\underline{a}$, the condition for event $\underline{e}$, t, the total number of stage-1 agents, n, the threshold number m, and the SPNE for the case that $0 < t-m$ and $t+m-2 < n$. Arrows to the right (left) indicate that, for this number of agents who take $\underline{a}$, any agent who takes $\overline{a}$ ($\underline{a}$) would want to deviate.

The case in which $t-m > 0$ and $n > t+m-2$ is presented in Figure 3.1. Intuitively, if less than $t-m$ agents take $\underline{a}$, the degree of causal responsibility for event $\underline{e}$ is low enough such that it is profitable to take $\underline{a}$. With every additional agent who takes $\underline{a}$ causal responsibility rises until, when $t-m$ agents take $\underline{a}$, any additional agent would increase causal responsibility high enough to deter taking $\underline{a}$. In the interval from $t-m$ to $t+m-2$, taking $\overline{a}$ gives a higher overall profit than taking $\underline{a}$. However, if more than $t+m-2$ agents take $\underline{a}$, causal responsibility is low enough again such that taking $\underline{a}$ is more profitable.

The following corollary describes the relationship of the threshold number m with the variables of the environment.

Corollary 1 (*Private gains* environment)**.**
The threshold number m is

i) increasing in the preference parameter ρ.

[23] The reason for $t-m$ in the case of the lower bound and $t+m-2$ in the case of the upper bound is that both yield the same distance from pivotality for event $\overline{e}$, $(t-1)-(t-m) = m-1$ and $(t+m-2)-(t-1) = m-1$.

ii) decreasing in the cost of punishment, c.

iii) increasing in the difference in judgments of the two events, $j(\overline{e}) - j(\underline{e})$.

iv) decreasing in the difference in stage-1 payoffs of the stage-1 agents, $\pi_i^I(\underline{a}) - \pi_i^I(\overline{a})$.

Proof. See Appendix 3.A.1.

These parameters thus influence the number of agents who take action $\underline{a}$ in the "$t-m$"-equilibrium. For example, the smaller is agent K's difference in judgment of the two events (reflected by $j(\overline{e})-j(\underline{e})$), the more agents will take $\underline{a}$. Intuitively, the less a manager cares about the success of a project compared to its failure, the more workers will allow themselves to shirk, but not enough to really make the project fail. The parameter also influence whether an equilibrium in which everyone takes $\underline{a}$ exists. If $t+m-2 \geq n$, then there never exists an equilibrium in which everyone takes $\underline{a}$ and $\underline{e}$ is implemented. Therefore, the existence depends on the difference $n-t$ and the variables that determine m.

In the *collective gains* environment, equilibrium predictions differ substantially. Specifically, there always exists an equilibrium in which all stage-1 agents take action $\overline{a}$ and event $\overline{e}$ is implemented. Furthermore, if $\underline{e}$ is implemented in equilibrium, it is only implemented by the smallest number of agents necessary.

Proposition 2 (*Collective gains* environment)**.**
(1) Suppose $\rho > c \cdot \frac{\pi_i^I(\underline{e})-\pi_i^I(\overline{e})}{j(\overline{e})-j(\underline{e})}$. Then there exists a unique subgame perfect Nash equilibrium in which all stage-1 agents choose action $\overline{a}$, event $\overline{e}$ is implemented, and agent K allocates according to equation 3.8.

(2) Suppose $\rho \leq c \cdot \frac{\pi_i^I(\underline{e})-\pi_i^I(\overline{e})}{j(\overline{e})-j(\underline{e})}$. Then there exist two subgame perfect Nash equilibria. One in which all stage-1 agents choose action $\overline{a}$, event $\overline{e}$ is implemented, and agent K allocates according to equation 3.8. And another one, in which t agents choose action $\underline{a}$, event $\underline{e}$ is implemented, and agent K allocates according to equation (3.8).
Proof. See Appendix 3.A.1.

The case in which condition (2) of Proposition 2 holds is shown in Figure 3.2. The intuition goes as follows: Assume all agents take $\underline{a}$ and event $\underline{e}$ is implemented. In this case, each agent has an incentive to deviate. By deviating he can still enjoy the higher stage-1 payoff while increasing his stage-2 payoff. Causal-responsibility attribution, in this case, induces the classic free-rider problem. Other theories do not do so as they don't predict differential allocation decisions as soon as more than t agents take action $\underline{a}$. This reasoning holds until exactly t agents take $\underline{a}$. These t agents are all fully causally

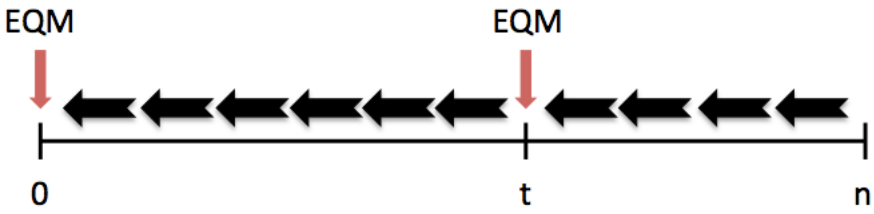


Figure 3.2: *Collective gains* environment: The line shows the number of stage-1 agents that choose $\underline{a}$, the condition for event $\underline{e}$, t, the total number of stage-1 agents, n, and the SPNE if part (2) of Proposition 2 holds. Arrows to the right (left) indicate that, for this number of agents who take $\underline{a}$, any agent who takes $\overline{a}$ ($\underline{a}$) would want to deviate.

responsible for event $\underline{e}$. Any additional agent who switches to taking $\overline{a}$ will induce a lower stage-1 payoff for himself and all other stage-1 agents. Therefore, if agent K's allocation decision is not deterrent in the case of full causal responsibility for event $\underline{e}$ (part (1) of Proposition 2), then an equilibrium exists in which exactly t agents take $\underline{a}$. If less than t agents take $\underline{a}$, then there exists again an incentive to deviate to $\overline{a}$ as doing so does not influence the payoff in stage-1 but increases the payoff in stage-2, such that, in a second equilibrium, all agents take $\overline{a}$.

To summarize, the analysis has shown that variations in the stage-1 environment can lead to very different equilibrium predictions under the causal responsibility framework. Two features stand out.

First, if $\underline{e}$ is implemented in equilibrium, then all n agents take action $\underline{a}$ in the *private gains* environment, while only the minimal necessary number, t, take $\underline{a}$ in the *collective gains* environment. Therefore, in equilibrium, causal responsibility is maximally diffused among all agents in the former environment, while it is fully focused on the t agents that take $\underline{a}$ in the latter environment. On the other hand, if $\overline{e}$ is implemented in equilibrium, then there might still be some agents who take $\underline{a}$ in the *private gains* environment but no agent takes $\underline{a}$ in the *collective gains* environment.

Second, the conditions for which $\underline{e}$ is implemented in equilibrium differ among the two environments. In the *collective gains* environment, $\underline{e}$ is only implemented in equilibrium, if $\rho \leq c \cdot \frac{\pi_i^I(\underline{e})-\pi_i^I(\overline{e})}{j(\overline{e})-j(\underline{e})}$, whereas, in the *private gains* environment, $\underline{e}$ is implemented in equilibrium, if $\rho \leq (1+n-t)\cdot c \cdot \frac{\pi_i^I(\underline{a})-\pi_i^I(\overline{a})}{j(\overline{e})-j(\underline{e})}$. Hence, if $\pi_i^I(\underline{e})-\pi_i^I(\overline{e}) = \pi_i^I(\underline{a})-\pi_i^I(\overline{a})$, then the condition for the existence of an equilibrium in which $\underline{e}$ is implemented is weaker in the *private gains* environment than in the *collective gains* environment (as $1+n-t>1$). This is because causal responsibility is fully diffused in the private gains environment, whereas diffusion of responsibility is prevented by the free-riding problem in the collective gains environment. Furthermore, the existence of the equilibrium depends on the total number of agents, n, in the *private gains* environment. Only if n is large enough is

causal responsibility for event $\underline{e}$ diffused enough to allow an equilibrium. The number for which this is the case can be interpreted as a critical mass of agents that results in such herding behavior. In the *collective gains* environment, on the other hand, the existence is independent of the total number of agents.

These equilibrium predictions seem to coincide with what is observed in reality. In *collective gain* environments, critical reforms are often implemented in parliament with only a small majority by the ruling party and the existence of cartels often crucially depend on each member. Even though all politicians might benefit from the implementation of a reform and all firms in an industry benefit from the higher prices caused by a cartel, blame and punishment is nevertheless only attributed to those who induced the events. In *private gain* environments, introspection suggests that the larger is the difference between n and t, i.e. the number of agents in the environment and the threshold level necessary to induce a certain event, the more likely it is that, in equilibrium, every agent takes the blameworthy action and the bad event is implemented. For example, flying is a widely enjoyed and few spend time thinking about their causal responsibility for the occurring emissions. This might be because the number of persons is far away enough from the number necessary for a plane to fly such that the equilibrium in which everyone takes advantage of flying exists. When people are fully causally responsible for emissions, for example, when driving a car, they are much more aware of their full causal responsibility and therefore hesitate more to cause the same amount of emissions.

3.4 Experimental evidence

While section 3.3.4 demonstrated that the theory of causal responsibility makes predictions distinctively different from prominent existing theories, such distinctiveness is certainly only desirable if it allows the explanation of data that cannot be explained by existing theories. While there exists no experimental study that was specifically designed to test the predictions of causal responsibility theory against other theories, in this section, I discuss experimental evidence from a range of existing studies. Each of these studies can be interpreted as a test of a separate aspect of the causal responsibility model. In the following, I focus on the different predictions for the allocation decision of agent K and abstain from full scale equilibrium analyses.

The key requirements to qualify for analysis are that the experiment i) consists of two stages: one stage in which potentially multiple agents and nature simultaneously choose actions that implement some stage-1 payoffs and another stage in which an agent independently decides about an allocation decision which influences the stage-2 payoffs, and ii) that actions and the associated payoffs are observable after stage 1. One challenge that

comes with interpreting existing experimental designs in light of responsibility theory is that the set of stage-1 events, X, is seldom explicitly modeled. However, in many cases, there are obvious candidates for such underlying events or it is possible to deduce them from the set of monetary stage-1 payoffs. For expositional ease, I assume that agent K faces a convex allocation cost function, which ensures an interior solution can exist.

3.4.1 The effect of moves of nature on causal responsibility

First, I will test the assumption that differences in ex post causal responsibility can be driven by nature's actions. To this end, I will turn to the study by Gurdal, Miller, and Rustichini (2013). In their experiment, a stage-1 agent A chooses between a risky and a safe asset, $A_A = \{r, s\}$, that determines the stage-1 payoff of an agent K. The safe asset gives agent K a certain payoff of π_c and the risky asset gives agent K a payoff of $\pi_h > \pi_c$, if nature chooses h, and a payoff of $\pi_l < \pi_c$, if nature chooses l. At the end of stage 1, agent K learns the choice of agent A, his payoff from stage 1 and, importantly, nature's choice, independent of whether agent A chose the safe or the risky asset. Agent A has a payoff of zero in stage 1. The experiment varies the payoffs of the two assets and the probability with which the lottery chooses h and l. Table 3.6 summarizes stage 1.

	h (p)	l $(1-p)$
r	π_h	π_l
s	π_c	π_c

Table 3.6: Stage-1 environment.

In the second stage, agent K has some money-equivalent points to distribute, at no cost to himself, between agent A and another agent B, who was not involved in stage-1. Agent K cannot keep the points for himself, but can refrain from distributing them. I denote the payoff that is given to agent A in stage 2 after history h as $p_A(h)$.

The authors find that the allocation to A is significantly higher when A chooses the risky asset and the lottery selects h than if he chooses the risky asset and the lottery selects l, thus $p_A(r, h) > p_A(r, l)$. This result cannot be explained by a theory of reciprocity, as the perceived kindness of A should only depends on his action and not on the action of the lottery. Furthermore, also outcome-based social preference theories have difficulty explaining the result, as agent K typically distributed all points. Thus, any allocation that decreased the payoff inequality between K and A, increased the payoff inequality between K and B.

Causal responsibility-based preferences, however, have no difficulty explaining this finding. First, there are three natural candidates for events in stage 1, namely the realizations of π_h, π_l, and π_c. For any reference payoff, the agent K judges the events in the order of the monetary payoff that they grant him, $j(\pi_h) > j(\pi_c) > j(\pi_l)$. Second, agent A always bears full ex post causal responsibility for the event that is implemented in stage 1 as he could have changed it by changing his action.[24] Thus, the theory predicts that the allocation to agent A is higher when A chooses r and the lottery chooses h than if A chooses r and the lottery chooses l, because agent K holds A responsible for a good and a bad result in stage-1.

The authors also find that $p_A(s,h) < p_A(s,l)$, i.e. agent K gave more to agent A when he chose the safe asset and the lottery chose l than when he chose the safe asset and the lottery chose h. However, the effect was much smaller in size and in many configurations not significant. Again, this result cannot be explained by traditional theories of reciprocity or inequity aversion. In this case, also a theory of responsibility preferences has difficulties explaining the result, because both, the judgment of the event as well as the causal responsibility level of agent A is identical for the two payoffs.[25]

3.4.2 The effects of causal responsibility for unrealized events

In section 3.4.1, an agent A who chooses the risky asset was fully ex post causally responsible for the actually implemented event and partially causally responsible for the hypothetical event that would have been implemented had the lottery chosen differently. However, this did not change the predictions for the allocation decision and was therefore neglected in the discussion. Now, I will turn to evidence that suggests that causal responsibility for a hypothetical event does, in fact, influence allocation decisions. This can be the case, for example, when a stage-1 agent bears full ex post causal responsibility for a realized good and partial ex post causal responsibility for an unrealized bad event. In this case, he will receive a lower allocation compared to an agent who bears full ex post causal responsibility for the good event only and the reason for this lies in the additive nature of the overall judgment of the agent's behavior. Furthermore, in the presence of a lottery, unrealized events can also matter for overall judgment of an agent's behavior through their impact on ex ante causal responsibility.

[24] Note that if A chooses r, he is also partially ex post causally responsible for the payoff that didn't realize. We will ignore this as well as the ex ante causal responsibility evaluation, for now, because it does not change the relevant predictions.

[25] However, an extended model of causal responsibility could not only incorporate causal responsibility for the *implementation* of an event, but also incorporate causal responsibility for the *prevention* of an event. Such an extension could explain why the allocation is lower after (s,h) - choosing the safe option makes agent A fully ex post causally responsible for the prevention of the best event - and higher after (s,l) - choosing the safe option makes agent A fully ex post causally responsible for the prevention of the worst event.

An experiment that allows to examine such situations is the study by Bartling, Engl, and Weber (2014). In the experiment, a dictator chooses between two possible actions, $A_D = \{u, d\}$, under two possible states of the world, $A_0 = \{l, r\}$. The state of the world is chosen by nature with equal probability and determines whether an action, which benefits the dictator, benefits or harms the recipient. While, in the baseline, the state of the world was known to the dictator, in the treatment, the dictator was initially ignorant about the state of the world, but could reveal it at no cost. Conforming to the responsibility framework, I interpret the state of the world as the action of nature. Table 3.7 summarizes the payoffs for the dictator and the recipient for the cases in which the dictator is informed or not informed about the state of the world. It is plausible to assume that the four possible payoff allocations between the dictator and the recipient mirror four possible events: an unfair (e_1), a fair (e_2), a dominant (e_3), and a dominated (e_4) event. A third party observes the actions of the dictator and the state of the world and decides whether and to what extent to punish the dictator.

	l
u	e_1 $(70, 10)$
d	e_2 $(50, 50)$

	r
u	e_3 $(70, 50)$
d	e_4 $(50, 10)$

	l (0.5)	r (0.5)
u	e_1 $(70, 10)$	e_3 $(70, 50)$
d	e_2 $(50, 50)$	e_4 $(50, 10)$

Table 3.7: Payoffs of the dictator and the recipient. Left and middle panel: Dictator is informed about the state of the world. Right panel: Dictator is not informed about the state of the world.

What are the predictions for allocations if the third party allocates according to the causal responsibility motive? First, under the assumption that the third party judges the events in the same way as the recipient, and with an interior reference payoff, he will judge event e_1 and e_4 as bad and event e_2 and e_3 as good, thus $j(e_2) > 0 > j(e_1)$ and $j(e_3) > 0 > j(e_4)$.[26] Second, he evaluates the dictator's causal responsibility for the respective events according to Table 3.8. If the dictator is informed about the state of the world, he is fully causally responsible for whatever event he implements as his action is always pivotal and ex ante and ex post causal responsibility coincide. For example, if nature chose l, the dictator knows this and chooses u, then he is fully causally responsible for event e_1. The comparative-statics predictions for the third party's allocation to the dictator in the case of an informed dictator would thus be $p_D(u|l) < p_D(d|l)$ and $p_D(u|r) > p_D(d|r)$, where

[26]The assumption that the third party has similar judgments as the recipient is reasonable, in this case, because the third party and the recipient were only told at the very end of the experiment which role they actually had. The dictator, on the other hand, always knew his role.

$p_D(a_D|a_0)$ denotes the allocation after the dictator takes a_D in the case that state a_0 was known to him. Note that a *lower* allocation is equivalent to a *higher* punishment.

Informed?	a_D	a_0	ω	r^{EP}_{D,e_1}	r^{EP}_{D,e_2}	r^{EP}_{D,e_3}	r^{EP}_{D,e_4}	r^{EA}_{D,e_1}	r^{EA}_{D,e_2}	r^{EA}_{D,e_3}	r^{EA}_{D,e_4}
yes	u	l	$(1,0)$	1	0	0	0	1	0	0	0
yes	d	l	$(1,0)$	0	1	0	0	0	1	0	0
yes	u	r	$(0,1)$	0	0	1	0	0	0	1	0
yes	d	r	$(0,1)$	0	0	0	1	0	0	0	1
no	u	l	$(\frac{1}{2},\frac{1}{2})$	1	0	$\frac{1}{2}$	0	$\frac{1}{2}$	0	$\frac{1}{2}$	0
no	d	l	$(\frac{1}{2},\frac{1}{2})$	0	1	0	$\frac{1}{2}$	0	$\frac{1}{2}$	0	$\frac{1}{2}$
no	u	r	$(\frac{1}{2},\frac{1}{2})$	$\frac{1}{2}$	0	1	0	$\frac{1}{2}$	0	$\frac{1}{2}$	0
no	d	r	$(\frac{1}{2},\frac{1}{2})$	0	$\frac{1}{2}$	0	1	0	$\frac{1}{2}$	0	$\frac{1}{2}$

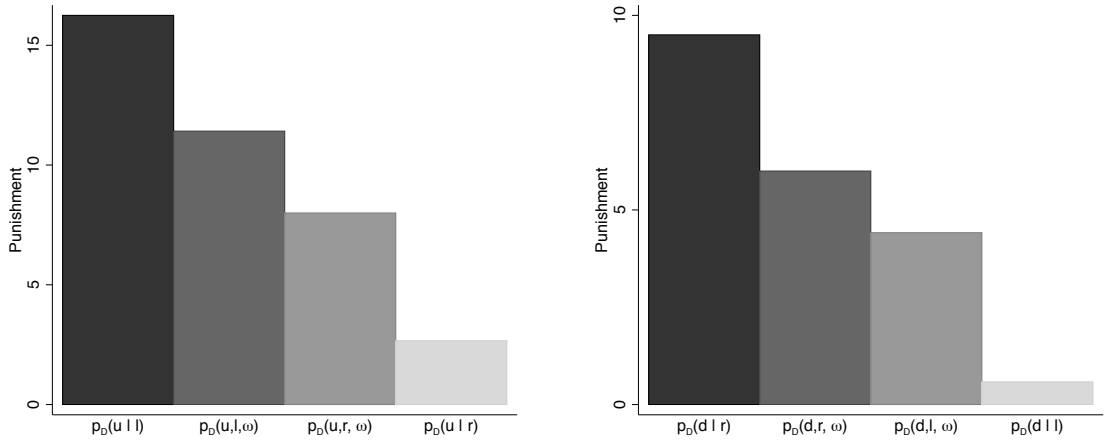
Table 3.8: Ex post and ex ante causal responsibility levels.

If the dictator is not informed about the state of the world, however, causal responsibility attribution is different. First, the dictator still bears full ex post causal responsibility for the event that is implemented, because he could always change the event by changing is action. In addition, however, the dictator also bears partial ex post causal responsibility for the hypothetical event that would have realized if nature would have chosen differently. For example, suppose an uninformed dictator chooses u and nature chooses l and thus the event e_1 realizes. The dictator bears full ex post causal responsibility for the bad event e_1, but he also bears partial ex post causal responsibility for the good event e_3 - to degree $\frac{1}{2}$ - that would have realized if the lottery would have chosen otherwise. Second, if the dictator is not informed, he will bear ex ante causal responsibility for both possible events that can follow from his action choice and the degree of ex ante causal responsibility is $\frac{1}{2}$ for both events as nature chooses with equal probability.

Suppose an uninformed dictator chooses u, nature chooses l and thus the bad event e_1 is implemented. Two effects influence the overall judgment of the dictator's behavior compared to the situation when he was informed. First, the dictator still bears full ex post causal responsibility for event e_1, but the fact that he also bears *lower* ex ante causal responsibility for e_1 reduces his overall causal responsibility for e_1 if agent K places some weight on ex ante causal responsibility. Second, the fact that the dictator additionally bears partial causal responsibility for the unrealized good event e_3 that would have realized if nature had chosen otherwise, furthermore increases the overall judgment of the dictators behavior. Note that, if the third party would place weight on ex ante causal responsibility only, then he would judge the overall behavior of the dictator independent of the choice of nature, and thus would also allocate independently. The complete allocation predictions are thus as follows:

Prediction 1. *The comparative statics predictions for allocations to the dictator are* $p(u|l) < p(u,l,(\frac{1}{2},\frac{1}{2})) \leq p(u,r,(\frac{1}{2},\frac{1}{2})) < p(u|r)$ *if the dictator chooses* u*, and* $p(d|r) <$

$p(d, r, (\frac{1}{2}, \frac{1}{2})) \leq p(d, l, (\frac{1}{2}, \frac{1}{2})) < p(d|l)$ *if the dictator chooses d.*



(a) Dictator chooses u

(b) Dictator chooses d

Figure 3.3: Observed punishment for dictator. Note that a *higher* punishment is equivalent to a *lower* allocation.

Figure 3.3 shows the observed punishment attributed to the dictator after he chooses u or d in both possible situations. The observed punishment pattern coincides with the allocation predictions (a higher punishment equals a lower allocation). Every predicted difference is statistically significant in the experimental data.[27] Interestingly, punishment still differs for an uninformed dictator, depending on the choice of nature. This can be interpreted as evidence that the third party does not place weight on ex ante causal responsibility only. Note that neither intention- nor outcome-based social preference theories predict a similar punishment pattern (see discussion in the paper).

Relatedly, in Bartling and Fischbacher (2012), a player A could either implement a payoff distribution that grants him and a player B a high, but player C a low payoff (unfair event), or a perfectly equal payoff distribution (fair event). Alternatively, player A could delegate the decision to player B, who then faced the identical choice. After observing the actions of player A and (potentially) player B, player C could distribute costly punishment points to player A and B. This sequential setting can be analyzed with the same tools as the simultaneous setting. Since nature is absent, ex ante and ex post causal responsibility coincide .

For any interior reference point, player C judges the unfair allocation as bad and the fair allocation as good. Punishment predictions are then as follows: Player C should punish a player A who implements the unfair event on his own as much as a player B who

[27]The figure shows the data from the treatment in which the dictator could reveal the state of the world, or remain uninformed. Using instead the data from the baseline for the allocation decisions of an informed dictator does not change the qualitative pattern and significance.

implements the unfair allocation after being delegated the decision. In both cases, the players bear full causal responsibility for the unfair allocation only, as they could have implemented the fair allocation instead. Indeed, Bartling and Fischbacher find positive punishment in both cases with no significant difference in punishment between the two cases.

On the other hand, if player A delegates, then causal responsibility theory predicts that he should be punished less if player B implements the unfair allocation than if he would have implemented it himself. In this case, player A is fully causally responsible for the bad unfair but also partly causally responsible for the good fair allocation that would have been implemented had player B chosen otherwise. Again, this prediction is confirmed by the data: player A is still punished, but significantly less than if he would have implemented the unfair allocation himself. However, if player B implements the fair allocation after delegation, then player A is not punished more than if he implements the fair allocation himself. According to causal responsibility, player A should receive a higher allocation in the second case. However, as the experiment only allowed payoff reductions (punishment), this could be an artifact of the design. Again, outcome-based and intention-based social preference theories cannot explain these results (see discussion in the paper).

3.4.3 The effect of partial causal responsibility

The last two sections analyzed situations in which a decision maker was fully ex post causally responsible for an implemented event and, sometimes, in addition, partially ex post causally responsible for a hypothetical event that was judged differently. In this section, I look at situations in which an agent has positive causal responsible for one event only, but his causal responsibility for that event varies from full to partial to zero. As the event, and thus the judgment, is held constant, a ceteris paribus change in causal responsibility for an event that is judged bad or good should lead to an according change in the allocation decision.

Unfortunately, I am not aware of experimental evidence with simultaneous decision making in stage 1 that would allow such a direct test of the theory. However, there exists experimental evidence with sequential decision making in stage 1, which can be interpreted "as if" agents acted simultaneously. In Bartling, Fischbacher, and Schudy (forthcoming), three agents, $I = \{A, B, C\}$, sequentially vote, $A_i = \{u, f\}$, on which of two possible payoff allocation between them and three other agents to implement in stage 1 . One is an unfair allocation, granting the three stage-1 agents a payoff of 9 and the other three agents a payoff of 1, and the other is a fair allocation, granting all agents a payoff of 5. After all three votes are cast, the unfair allocation is implemented, if at least two

agents voted u, otherwise the fair allocation is implemented. The set of events can thus be described as $X = \{unfair, fair\}$ and the voting rule provides the function that links actions to events. This environment is therefore a sequential version of the *collective gain* environment discussed before.

After observing the actions of all three agents and the stage-1 payoff, the three other agents independently attribute costly punishment points to each of the three stage-1 agents.[28] Since only the allocation of one of the three stage-2 agents was randomly selected and implemented, each of the three stage-2 agents can be interpreted as an independent agent K. Furthermore, since the payoff of agent K from the unfair event is lower than from the fair event, agent K judges the unfair (fair) allocation as a bad (good) event, $j(fair) > 0 > j(unfair)$, independent of the reference payoff. Agent K could only reduce the stage-1 agents' payoffs and only the choice of no punishment was costfree.

While voting happens sequentially, one can still analyze causal responsibility "as if" it would happen simultaneously. For example, if all three stage-1 agents vote for the unfair allocation, the vote of each agent was not pivotal for its implementation, but would have been pivotal, if one of of the other agents would have voted differently. Hence, for this action profile, each stage-1 agent has partial causal responsibility for the unfair allocation.[29] How would the predictions for punishment look like, if we interpret the experiment in such a way? First, since nature is not present, ex ante and ex post causal responsibility coincide. Second, in this environment, agents can only be causally responsible for one event, either the unfair one, if they voted u, or the fair one, if they voted f. Since only punishment was allowed, I henceforth focus on causal responsibility for the unfair event. Three levels of causal responsibility are possible. First, an agent bears no causal responsibility for the unfair event, if he voted for f. Second, an agent bears full causal responsibility for the unfair event, if he voted for u and, in total, two agents voted for u and the third one for f. Third, an agent is partially causally responsible for the unfair event (with degree 1/2), if he voted for u and either no other agent, or two other agents voted for u. The comparative-statics predictions for punishment are therefore straightforward:

Prediction 2. *Punishment for a stage-1 agent is highest after histories in which he bears full causal responsibility for the unfair event, lower after histories in which he bears partial causal responsibility and lowest after histories in which he bears no causal responsibility.*

[28] The strategy method was used to elicit punishment decision, allowing a within-subject test of differences in punishment.

[29] Of course, treating the sequential vote "as if" it happened simultaneously might mask some important features of sequential decision making. For example, if A and B already voted for the unfair event, agent C knows that his vote has no influence anymore and this might influence how his decision is judged. On the other hand, the same argument could be made if the game were played simultaneously and all three stage-1 agents vote for unfair in a pure-strategy equilibrium. Also then, agent C's vote has no impact and everyone knows it.

Note that, in the case of partial causal responsibility, the theory predicts no difference in punishment between the case in which the unfair allocation is actually implemented (three u votes) and when it is not (one u vote). Thus, the design also allows a test of the hypothesis that causal responsibility for an hypothetical, unrealized event is judged in the same way as causal responsibility for a realized event.

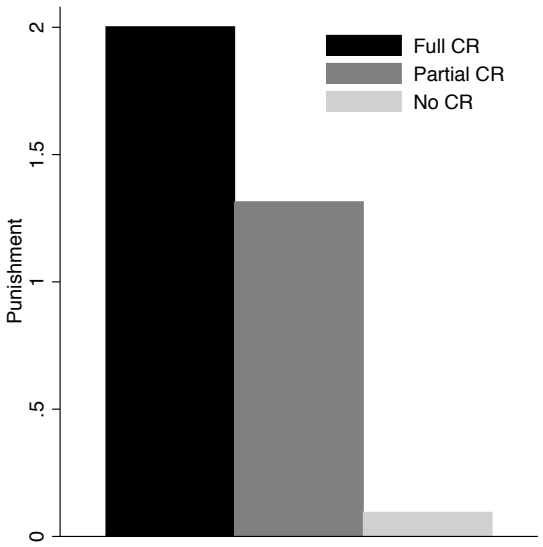


Figure 3.4: Average punishment depending on causal responsibility for the unfair event.

Figure 3.4 pools the data for all three stage-1 agents and shows the average punishment levels for the stage-1 agents depending on their causal responsibility for the unfair event. For example, the "Full CR"-bar shows the average punishment for stage-1 agents after any history in which they are fully causally responsible for the unfair event. The data clearly confirms the comparative-statics predictions. A subject who is fully causally responsible for the unfair event is punished significantly more than a subject who is partially causally responsible (2.00 vs. 1.31; Wilcoxon signed-rank test, p=0.000). Similarly, partial causal responsibility for the unfair event is punished significantly more than no causal responsibility (1.31 vs. 0.09; Wilcoxon signed-rank test, p=0.000).[30] Furthermore, the data also confirms the prediction that causal responsibility for an event does not depend on the event's actual implementation. An agent who is partially causally responsible for the unfair event is not punished differently depending on whether the unfair event is actually implemented or not (1.40 vs. 1.23; Wilcoxon signed-rank test, p=0.670).

Importantly, the qualitative punishment pattern is still found when looking at the three stage-1 agents separately, which indicates that the sequential nature of the voting decision does not influence the qualitative punishment pattern predicted by causal responsibility. For each of the three stage-1 agents, punishment is always highest in case of full causal responsibility and lowest in case of no causal responsibility for the unfair event. The

[30]I am grateful to the authors of the study for letting me analyze their data.

differences between the three levels are statistically significant with the exception of agent B for whom the difference in punishment between full and partial causal responsibility is not significant (Wilcoxon signed-rank test, p=0.133). Furthermore, for all three agents, punishment for partial causal responsibility does never differ significantly depending on whether the unfair event is actually implemented, or not.

Table 3.9: Robustness of causal responsibility as punishment motive

	(1) OLS Pun.	(2) OLS Pun.	(3) OLS Pun.	(4 OLS) Pun.	(5) OLS Pun.	(6) OLS Pun.	(7) OLS Pun.
Causal responsibility	1.956*** (0.193)	1.903*** (0.202)	1.375*** (0.305)	1.215*** (0.194)	1.606*** (0.248)	1.737*** (0.192)	0.657*** (0.208)
Outcome unfair		0.073 (0.113)					0.048 (0.070)
Choice unfair			0.532** (0.231)				0.453*** (0.161)
"Intention unkind"				0.719*** (0.157)			0.517** (0.197)
Choice unfair X "Intention unkind"					0.360 (0.216)		0.042 (0.227)
"Pivotality"						0.452*** (0.161)	0.403** (0.155)
Constant	0.143*** (0.041)	0.127** (0.057)	0.095*** (0.033)	0.122*** (0.041)	0.154*** (0.038)	0.150*** (0.041)	0.083** (0.037)
Observations	1728	1728	1728	1728	1728	1728	1728
Adjusted R^2	0.262	0.262	0.267	0.274	0.265	0.269	0.281

Notes: The dependent variable is attributed punishment points for voters. Besides the causal responsibility variable, the other explanatory variables are constructed as in Bartling, Fischbacher, and Schudy (forthcoming): *Outcome unfair* is a dummy variable which equals 1 if the unfair allocation is implemented. *Choice unfair* is a dummy variable which equals 1 if the $a_i = u$ is chosen. *"Intention unkind"* is a dummy variable equal to 1 if the respective voter opted for the unfair allocation and no majority was achieved before her vote. *"Pivotality"* is a dummy equal to 1 if the $a_i = u$ is chosen, the unfair allocation occurred and the respective voter was the second voter opting for the unfair allocation. Robust standard errors (clustered on 72 individuals) in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

When regressing the level of causal responsibility on punishment (see column (1) of Table 3.9), I find a large positive and highly significant effect. Furthermore, the adjusted R^2 of 0.262 shows that causal responsibility can explain a large part of the variation in punishment. However, it might be that omitted variables both influence the causal responsibility measure as well as punishment. For example, a positive causal responsibility level is only

possible when the agent voted for the unfair event and voting for an unfair event might itself be an independent motive for punishment. Hence, not including the voting decision in the regression can lead to an omitted variable bias in the estimation of the causal responsibility coefficient. In order to check the robustness of the effect of causal responsibility on punishment, I include, separately and jointly, all punishment motives considered in the paper of Bartling et al. (see column (2)-(7) of Table 3.9). Causal responsibility remains highly significant in all regressions. Besides causal responsibility, voting for the unfair allocation, "unkind intentions" and "pivotality" (both in the definition of Bartling et al.) have significant and positive effects on punishment. Whether the outcome is unfair or fair has no significant effect on punishment.[31]

To summarize, the past three sections lent support to the assumptions that i) actions of lotteries matter for the evaluation of causal responsibility, ii) causal responsibility can not only be explained by the ex ante component and causal responsibility for unrealized events matters, and iii) partial causal responsibility is predictive of allocation decisions. For each of the three parts, the theory of causal responsibility makes better predictions than the most established competing theories. What is still missing is a test of partial ex post causal responsibility in a real simultaneous-move environment and a test of ex ante causal responsibility, i.e. a test in an environment with a lottery that varies the probability distribution of the lottery while holding everything else constant.

3.5 Discussion

This sections discusses some limitations of the framework and lies out directions for future research. Specifically, possible generalizations of the causal responsibility function, generalizations of the set of agents for whom causal responsibility is evaluated, and the possible influence of psychological factors for responsibility perceptions are considered.

First, note that the distance to pivotality measure in equation 3.1, simply counts the minimum number of changes necessary to achieve pivotality, thereby treating every change of action equally. One might argue that some changes are more likely than others, for example, because they are cheaper, and that therefore an agent whose distance to pivotality is

[31] Bartling et al. define a stage-1 agents vote as "pivotal", if the respective agent was the second voter opting for the unfair allocation. I, on the other hand, view an agent's vote as pivotal for an event, if the event counterfactually depends on the vote. For example, if agent A and B vote u and agent C votes f, I define both agent A and B's votes as pivotal for the unfair allocation, whereas Bartling et al. would only count agent B's vote as pivotal. Controlling for pivotality in my definition still leads to a highly significant effect of causal responsibility, suggesting that it is not only predictive because it incorporates pivotality (i.e. full causal responsibility), but also because partial causal responsibility is important. However, it is not possible to include pivotality in my definition in the full regression as this would lead to collinearity (the dummy for the vote choice and the dummy for pivotality together perfectly explain all variation in the causal responsibility variable).

three "likely" changes is more causally responsible than another for whom the distance is three "unlikely" changes. An appropriate weighting of action changes in the function that counts the differences between two action profiles, $c(y, x)$, could accommodate such a feature. The distance function in equation 3.1 would then minimize not the absolute number of changes, but the weighted number of changes.

Relatedly, I have only discussed discrete (and often binary) action sets, but the notion of causal responsibility could be extended to continuous action sets. In many cases, agents don't decide about using clean or dirty technology, but about with what intensity to use a certain technology. For example, governments do not only decide whether they want to allow the emission of greenhouse gases or not, but about how much emissions they permit. In this case, one might want to evaluate a country's causal responsibility for the failure or success of the world to stay below the target of a temperature increase of maximally 2 degrees Celsius that was set in the Copenhagen Agreement.[32] Suppose temperatures rise by more than 2 degrees Celsius if more than 100 units of greenhouse gases are emitted. Furthermore, suppose the US and China each emit 55 units, Germany emits 20 units and Fiji emits 1 unit, such that, in total, 131 units are emitted and the target is missed. An appropriate measure of causal responsibility in this case should not only take the number of changes necessary to reach pivotality into account (0 for the US and China and 1 for Germany and Fiji), but also apply an appropriate weighting of the emission levels.

Second, instead of formulating a theory that explains how people allocate blame and praise to other people depending on those *other* peoples' causal responsibility, one could just as well formulate a theory in which people care about their *own* causal responsibility for an event.[33] In fact, there exists experimental evidence suggesting that such a theory can successfully predict how people choose in situations in which there is a trade-off between benefits to oneself and the violation of some ethical standard. For example, in the experiment of Falk and Szech (2013), subjects had to choose between killing a mouse and receiving some money (option B) and saving the mouse and receiving no money (option A) in the baseline condition.[34] Thus, in this condition, every subject was fully causally responsible (ex post and ex ante) for saving or killing the mouse and the theory suggests that subjects would only kill the mouse if the monetary compensation was worth more to them than the intrinsic blame resulting from bearing full causal responsibility for

[32] `http://unfccc.int/files/meetings/cop_15/application/pdf/cop15_cph_auv.pdf`

[33] For example, a utility function of the form $u_i(\mathbf{a}, \omega) = \pi_i(\mathbf{a}) + \rho_i \sum_{x \in X} r_{i,x}(\mathbf{a}, \omega) \cdot j_i(x)$ incorporates feelings of blame and praise towards oneself for the causal responsibility for the implementation of some event x.

[34] Relatedly, Charness (2000) study experimentally how the presence of an external party or a lottery can alter a worker's feeling of responsibility and, as a result, change his effort provision and Charness and Jackson (2009) study how a subject's investment decision changes once it also influences and thus is responsible for another subject. Other studies experimentally examine the role of organizational hierarchies Ellman and Pezanis-Christou (2010) or the option to delegate decisions Hamman, Loewenstein, and Weber (2010) on outcomes and refer to responsibility motives in explaining their results.

the death of a mouse.

In the "diffused pivotality" treatment, subjects were matched in groups of eight and again given the choice between not receiving money (option A) and receiving some money (option B). The key change is that if at least one out of the eight subjects chose option B, eight mice would be killed, irrespective of the choices of the other subjects. They find that, compared to the baseline, a significantly larger proportion, 58.6 percent, of subjects chose option B in this treatment. Furthermore, they find that the probability to choose option B decreases monotonically with ones (non-incentivized) belief of being pivotal for the death of the mice, i.e. being the only one in the group that chooses B. Both findings are very much in line with what a theory of intrinsic causal responsibility attribution would predict. Namely, the more subjects choose option B, the lower is each of their causal responsibility for the death of the mice. Hence, if a subject believes that other subjects will choose option B, he is more likely to choose option B himself even if he would not have chosen option B in the baseline treatment. This also exemplifies how few persons who care only about their material payoff can trigger a cascade of causal responsibility diffusion after which even persons who care a lot about the negative externalities of their actions take these actions because their causal responsibility is diffused enough to not deter them anymore.

Third, I have so far only considered the causal responsibility for an event of agents whose actions directly feed into the event's "production" function, f, which can be interpreted as "direct" causal responsibility. I call these agents "first-order" agents. In addition, one could extend the notion to also evaluate causal responsibility for "second-order" agents whose actions do not directly feed into the function f, but whose actions influence the actions of the "first-order" agents and, by doing so, indirectly influence the event. For example, if a consumer demands a cheap product and the production of a cheap product is only possible if the firm exploits its workers, then the consumer's action does not directly determine whether workers are exploited, but the firm's decision will be influenced by the demand of the consumer. The current version would not attribute any causal responsibility to the consumer as he could never be directly pivotal for the event. Hence, it makes sense to attribute "first-order" causal responsibility to the firm, but also "second-order" causal responsibility to the consumer. Such "higher-order" causal responsibility can be evaluated by the same methods as before. For example, given a firm's best-response function, which is rationally anticipated, a consumer's "second-order" causal responsibility for the exploitation of the workers could be equal to the sum of the distance of his action from being pivotal for the firm's action and the distance of the firm's action from being pivotal for the exploitation of the workers.

Fourth, a considerable body of evidence in moral psychology suggests that in addition to rational reasoning, also emotional and intuitive factors play a role in moral decision making

(Haidt, 2001; Haidt and Kesebir, 2009; Cushman, Young, and Greene, 2010; Greene, 2013). If this is the case, emotional factors are also likely to influence perceptions of responsibility. Hence, one should not interpret causal responsibility as the single relevant factor, but as a "rational" benchmark, against which other factors can be tested. For example, there exists evidence that people perceive the implementation of an "bad" event as less bad, when it was implemented due to an omission (not changing the status quo) as opposed to when it was implemented due a commission (changing the status quo) Cox, Servatka, and Vadovic (2013). Furthermore, increased spatial, temporal and social distance might reduce perceptions of responsibility, as might whether an event was implemented as a means or as a side-effect (for a discussion of these channels, see Greene (2013)).

All of these points highlight that there is considerable potential for future work and that the present study should be seen as a starting point for the theoretic and empirical analysis of (causal) responsibility perceptions in economic contexts.

3.6 Conclusion

This paper introduces the notion of causal responsibility into the economic framework. Causal responsibility measures the causal impact of an agent's action on the occurrence of an event. When taking causal responsibility into account, an agent who likes or dislikes the event will attribute reward or punishment to the agents involved in its implementation in relation to their causal responsibility for the event. In many group processes causal responsibility driven allocation decisions are distinctively different from what prominent existing theories would predict. In particular, causal responsibility makes better predictions when multiple actions determine the occurrence of an event, such that it is possible that no single action is pivotal for the event. In these cases, causal responsibility depends on an agent's action's distance from pivotality. By applying causal responsibility to two common environments, the paper demonstrates that, in equilibrium, causal responsibility for an event is maximally diffused between all, and, in other cases, maximally focused on some of the involved agents. In the former case, whether an event is implemented in equilibrium crucially depends on the number of active agents which determines the potential diffusion of causal responsibility. Finally, the paper tests the predictive power of the causal responsibility notion for allocation decisions in data from existing, incentivized experiments and finds that the causal responsibility motive can explain observed punishment patterns successfully and that it remains a highly significant predictor for punishment even after controlling for several other potential punishment motives. Several directions for promising extensions and applications of the notion of causal responsibility are provided.

To conclude, I want to discuss the relevance of causal responsibility-based preferences in markets. First, perceptions of causal responsibility and the associated attribution of blame can influence the willingness to consume goods that come with a negative externality. For example, many forms of long-distance transportation come with environmentally damaging emission of CO2. When facing the decision whether to use a certain transportation method, the consumer also faces a decision about how causally responsible he wants to be for the generated externality. Driving a car alone, for example, leads to full causal responsibility of the driver for the negative externality generated, as, if he would not use the car, no one else would. And, indeed, many people nowadays prefer not to use cars and the reason brought forward is often their negative environmental impact. On the other hand, when taking the plane, the same people seem much less concerned about their responsibility for the negative externality generated, even though emissions per passenger are often worse than if the same distance would be driven by car. Of course, the plane would also not fly without passengers and thus there exists a necessary number of passengers for the plane to fly. However, on a typical flight the number of passengers exceeds the number of necessary passengers and causal responsibility is therefore diffused among them. This diffusion can lead to a greater willingness to take planes compared to the situation in which each passenger would bear full causal responsibility for the negative externality.

Second, the evaluation of causal responsibility can also be linked to replaceability-arguments which are often made in markets. For example, a government might excuse his role in an arms deal with an authoritarian regime by stating that, if they would not have sold the arms, some other government would have, implicitly referring to a non-pivotality condition. The causal responsibility framework can capture such arguments by modeling each government as an independent agent who decides about whether to offer arms or not. Interestingly, it follows that the market structure is related to causal responsibility. A monopolist in a certain industry is pivotal for which goods are traded. For example, if a national monopolist on energy decides not to supply nuclear energy, it will not be supplied. This can explain why people often blame power companies, but not as much the individual electricity consumer for the existence of nuclear power plants and their externalities. The more firms have the ability to supply a certain good, the greater is the chance that there exists an equilibrium in which they all do it. One could therefore argue that, the number of firms operating in industries with sensitive goods should be controlled in order to focus causal responsibility on a few players.

Bibliography

Abramowitz, A. I., D. J. Lanoue, and S. Ramesh (1988): "Economic Conditions, Causal Attributions, and Political Evaluations in the 1984 Presidential Election," *The Journal of Politics*, 50(4), 848–863.

Alicke, M. (1992): "Culpable Causation," *Journal of Personality and Social Psychology*, 63(3), 368–378.

Alicke, M. D. (2000): "Culpable control and the psychology of blame.," *Psychological bulletin*, 126(4), 556–74.

Alicke, M. D., J. Buckingham, E. Zell, and T. Davis (2008): "Culpable control and counterfactual reasoning in the psychology of blame.," *Personality & social psychology bulletin*, 34(10), 1371–81.

Bartling, B., F. Engl, and R. A. Weber (2014): "Does willful ignorance deflect punishment? – An experimental study," *European Economic Review*, 70, 512–524.

Bartling, B., and U. Fischbacher (2012): "Shifting the blame: on delegation and responsibility," *Review of Economic Studies*, 79(1), 67–87.

Bartling, B., U. Fischbacher, and S. Schudy (forthcoming): "Pivotality and Responsibility Attribution in Sequential Voting," *Journal of Public Economics*.

Battigalli, P., and M. Dufwenberg (2009): "Dynamic psychological games," *Journal of Economic Theory*, 144(1), 1–35.

Beebee, H., C. Hitchcock, and P. Menzies (eds.) (2012): *The Oxford Handbook of Causation*. Oxford University Press.

Berg, S. V. (1982): "Causal responsibility and peak load pricing," *Energy Economics*, 4(4), 246–250.

Blanco, M., B. Çelen, and A. Schotter (2013): "On Blame and Reciprocity: An Experimental Study," *Working Paper*, pp. 1–34.

Bolton, G. E., and A. Ockenfels (2000): "ERC: A Theory of Equity, Reciprocity, and Competition," *American Economic Review*, 90(1), 166–193.

Boyd, R., H. Gintis, S. Bowles, and P. J. Richerson (2003): "The evolution of altruistic punishment.," *Proceedings of the National Academy of Sciences of the United States of America*, 100(6), 3531–5.

Bunzl, M. (1979): "Causal Overdetermination," *The Journal of Philosophy*, 76(3), 134–150.

Charness, G. (2000): "Responsibility and effort in an experimental labor market," *Journal of Economic Behavior & Organization*, 42(3), 375–384.

Charness, G., and M. O. Jackson (2009): "The role of responsibility in strategic risk-taking," *Journal of Economic Behavior & Organization*, 69(3), 241–247.

Chockler, H., and J. Halpern (2004): "Responsibility and Blame: A Structural-Model Approach," *J. Artif. Intell. Res.(JAIR)*, 22, 93–115.

Cox, J. C., M. Servatka, and R. Vadovic (2013): "Status Quo Effects in Fairness Games: Reciprocal Responses to Acts of Commission vs. Acts of Omission," *Working Paper*.

Cushman, F. (2008): "Crime and punishment: distinguishing the roles of causal and intentional analyses in moral judgment.," *Cognition*, 108(2), 353–80.

Cushman, F., L. Young, and J. D. Greene (2010): "Our multi-system moral psychology: Towards a consensus view," in *Oxford Handbook of Moral Psychology*, ed. by J. Doris et al., pp. 1–20. Oxford University Press.

Darley, J My, J., and B. Latané (1968): "Bystander Intervention in Emergencies: Diffusion of Responsibility," *Journal of Personality and Social Psychology*, 8(4), 377–383.

Darley, J. M., and T. R. Shultz (1990): "Moral Rules: Their Content and Acquisition," *Annual Review of Psychology*, 41(1), 525–556.

Dietrich, F. (2002): "Causal Responsibility and Rationing in Medicine," *Ethical Theory and Moral Practice*, 5, 113–131.

Duch, R., W. Przepiorka, and R. Stevenson (2014): "Responsibility attribution for collective decision makers," *American Journal of Political Science*, 59(2), 372–389.

Dufwenberg, M., and G. Kirchsteiger (2004): "A theory of sequential reciprocity," *Games and Economic Behavior*, 47(2), 268–298.

Ellman, M., and P. Pezanis-Christou (2010): "Organizational structure, communication, and group ethics," *American Economic Review*, 100(5), 2478–2491.

Falk, A., and N. Szech (2013): "Organizations , Diffused Pivotality and Immoral Outcomes," *Working Paper*, (May), 1–15.

Fehr, E., and U. Fischbacher (2004): "Third-party punishment and social norms," *Evolution and Human Behavior*, 25(2), 63–87.

Fehr, E., and S. Gachter (2002): "Altruistic punishment in humans," *Nature*, 415(6868), 137–140.

Fehr, E., and K. M. Schmidt (1999): "A Theory of Fairness, Competition, and Cooperation," *The Quarterly Journal of Economics*, 114(3), 817–868.

Gerstenberg, T., and D. a. Lagnado (2010): "Spreading the blame: The allocation of responsibility amongst multiple agents.," *Cognition*, 115(1), 166–71.

Goldman, A. I. (1999): "Why Citizens Should Vote: A Causal Responsibility Approach," *Social Philosophy and Policy*, 16(02), 201–217.

Gomez, B. T., and J. M. Wilson (2003): "Causal Attribution and Economic Voting in American Congressional Elections," *Political Research Quarterly*, 56(3), 271–282.

Greene, J. (2013): *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*. Penguin Press, New York.

Gurdal, M. Y., J. B. Miller, and A. Rustichini (2013): "Why Blame?," *Journal of Political Economy*, 121(6), 1205–1247.

Haidt, J. (2001): "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment," *Psychological review*, 108(4), 814–834.

Haidt, J., and S. Kesebir (2009): "Morality," *Handbook of Social Psychology*, pp. 1–61.

Halpern, J. Y., and J. Pearl (2005): "Causes and Explanations: A Structural-Model Approach. Part I: Causes," *The British Journal for the Philosophy of Science*, 56(4), 843–887.

Hamman, J. R., G. Loewenstein, and R. A. Weber (2010): "Self-Interest through Delegation: An Additional Rationale for the Principal-Agent Relationship," *American Economic Review*, 100(4), 1826–1846.

Hart, H. L. A. (1968): *Punishment and Responsibility*. Oxford University Press, 1st edn.

Hart, H. L. A., and A. Honore (1985): *Causation in the Law*. Oxford University Press, 2nd edn.

Heider, F. (1958): *The Psychology of Interpersonal Relations*. Wiley, New York, NY.

Iyengar, S. (1996): "Framing Responsibility for Political Issues," *Annals of the American Academy of Political and Social Science*, 546, 59–70.

Kahneman, D., and A. Tversky (1982): *Judgment Under Uncertainty: Heuristics and Biases*chap. The simulation heuristic. Cambridge University Press.

Kahneman, D., and C. a. Varey (1990): "Propensities and counterfactuals: The loser that almost won.," *Journal of Personality and Social Psychology*, 59(6), 1101–1110.

Konow, J. (1996): "A positive theory of economic fairness," *Journal of Economic Behavior & Organization*, 31(1), 13–35.

——— (2000): "Fair Shares: Accountability and Cognitive Dissonance in Allocation Decisions," *American Economic Review*, 90(4), 1072–1092.

——— (2001): "Fair and square: the four sides of distributive justice," *Journal of Economic Behavior & Organization*, 46(2), 137–164.

Lagnado, D. a., and S. Channon (2008): "Judgments of cause and blame: the effects of intentionality and foreseeability.," *Cognition*, 108(3), 754–70.

Lagnado, D. a., T. Gerstenberg, and R. Zultan (2013): "Causal responsibility and counterfactuals.," *Cognitive science*, 37(6), 1036–73.

Latané, B., and J. M. Darley (1968): "Group inhibition of bystander intervention in emergencies.," *Journal of personality and social psychology*, 10(3), 215–21.

Leibbrandt, A., and R. López-Pérez (2012): "An exploration of third and second party punishment in ten simple games," *Journal of Economic Behavior & Organization*, 84(3), 753–766.

Malle, B. F., S. Guglielmo, and A. E. Monroe (2014): "A Theory of Blame," *Psychological Inquiry*, 25(2), 147–186.

Manove, M. (1997): "Job responsibility, pay and promotion," *The Economic Journal*, 107(440), 85–103.

Miller, D. (2001): "Distributing Responsibilities," *Journal of Political Philosophy*, 9(4), 453.

Moore, M. S. (2009): *Causation and responsibility: An essay in law, morals, and metaphysics*. Oxford University Press.

Nikiforakis, N., and H. Mitchell (2013): "Mixing the carrots with the sticks: third party punishment and reward," *Experimental Economics*, 17(1), 1–23.

Pearl, J. (2000): *Causality: Models, Reasoning and Inference*, vol. 29. Cambridge University Press.

Prendergast, C. J. (1995): "A Theory of Responsibility in Organizations," *Journal of Labor Economics*, 13(3), 387.

Rabin, M. (1993): "Incorporating Fairness into Game Theory and Economics," *American Economic Review*, 83(5), 1281–1302.

Roese, N. J. (1997): "Counterfactual thinking.," *Psychological bulletin*, 121(1), 133–48.

Schlenker, B. R., T. W. Britt, J. Pennington, R. Murphy, and K. Doherty (1994): "The triangle model of responsibility.," *Psychological review*, 101(4), 632–52.

Sebald, A. (2010): "Attribution and reciprocity," *Games and Economic Behavior*, 68(1), 339–352.

Shaver, K. G. (1985): *The Attribution of Blame: Causality, Responsibility, and Blameworthiness*. Springer.

Sliwka, D. (2006): "On the notion of responsibility in organizations," *Journal of Law, Economics, and Organization*, 22(2), 523–547.

Sloman, S. a., P. M. Fernbach, and S. Ewing (2009): *Chapter 1 Causal Models: The Representational Infrastructure for Moral Judgment*, vol. 50. Elsevier Inc.

Sober, E. (1988): "Apportioning Causal Responsibility," *The Journal of philosophy*, 85(6), 303–318.

Spellman, B. (1997): "Crediting causality.," *Journal of Experimental Psychology: General*, 126(4), 323–348.

Spellman, B. a., and D. R. Mandel (1999): "When Possibility Informs Reality: Counterfactual Thinking as a Cue to Causality," *Current Directions in Psychological Science*, 8(4), 120–123.

Thompson, D. (1980): "Moral Responsibility of Public Officials: The Problem of Many Hands," *The American Political Science Review*, 74(4), 905–916.

Weiner, B. (1995): *Judgments of Responsibility: A Foundation for a Theory of Social Conduct*. The Guilford Press.

Woodward, J. (2003): *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.

Wright, R. W. (1985): "Causation in Tort Law," *California Law Review*, 73(6), 1735–1828.

Wright, R. W. (1988): "Causation, Responsibility, Risk, Probability, Naked Statistics, and Proof: Pruning the Bramble Bush by Clarifying the Concepts," *Iowa Law Review*, 73, 1001–1077.

Zultan, R., T. Gerstenberg, and D. a. Lagnado (2012): "Finding fault: causality and counterfactuals in group attributions.," *Cognition*, 125(3), 429–40.

3.A Appendix

3.A.1 Proofs

Theorem 1. *The following two statements are equivalent:*

a) P.1 and P.2 hold.

b) $r_{i,x}^{EP}(\mathbf{a}) = \frac{1}{d_{i,x}(\mathbf{a})}$.

Proof of Theorem 1.
a) $\rightarrow$ *b)*
Suppose P.1 and P.2 hold. I show by contradiction that in this case $r_{i,x}^{EP}(\mathbf{a}) = \frac{1}{d_{i,x}(\mathbf{a})}$ has to hold.

First, let $d_{i,x}(\mathbf{a}) = 1$. Then any $r_{i,x}^{EP}(\mathbf{a}) \neq 1$ would contradict P.1. Hence, $r_{i,x}^{EP}(\mathbf{a}) = 1$ if $d_{i,x}(\mathbf{a}) = 1$.

Second, let $d_{i,x}(\mathbf{a}) = 1$ and $d_{j,x}(\mathbf{a}) = n$ with $n \in \{2, 3, ..., |N|\}$.
Suppose $r_{j,x}^{EP}(\mathbf{a}) > \frac{1}{n}$. Then $\frac{r_{i,x}^{EP}(\mathbf{a})}{r_{j,x}^{EP}(\mathbf{a})} < n$. This contradicts P.2 which states that $\frac{r_{i,x}^{EP}(\mathbf{a})}{r_{j,x}^{EP}(\mathbf{a})} = \frac{d_{j,x}(\mathbf{a})}{d_{i,x}(\mathbf{a})} = \frac{n}{1} = n$.
Next, suppose $r_{j,x}^{EP}(\mathbf{a}) < \frac{1}{n}$. Then $\frac{r_{i,x}^{EP}(\mathbf{a})}{r_{j,x}^{EP}(\mathbf{a})} > n$. This contradicts P.2 which states that $\frac{r_{i,x}^{EP}(\mathbf{a})}{r_{j,x}^{EP}(\mathbf{a})} = \frac{d_{j,x}(\mathbf{a})}{d_{i,x}(\mathbf{a})} = \frac{n}{1} = n$. Hence, $r_{j,x}^{EP}(\mathbf{a}) = \frac{1}{n}$ if $d_{j,x}(\mathbf{a}) = n$.

Third, let $d_{i,x}(\mathbf{a}) = \infty$. Then any $r_{i,x}^{EP}(\mathbf{a}) \neq 0$ would contradict P.1. Hence, $r_{i,x}^{EP}(\mathbf{a}) = 0$ if $d_{i,x}(\mathbf{a}) = \infty$.

Taken together, it was shown that, if P.1 and P.2 hold, then

$$r_{i,x}^{EP}(\mathbf{a}) = \begin{cases} 1 & \text{if} \quad d_{i,x}(\mathbf{a}) = 1 \\ \dfrac{1}{n} & \text{if} \quad d_{i,x}(\mathbf{a}) = n \quad \text{with} \quad n \in \{2, 3, ..., |N|\} \\ 0 & \text{if} \quad d_{i,x}(\mathbf{a}) = \infty. \end{cases}$$

Which is mathematically equivalent to $r_{i,x}^{EP}(\mathbf{a}) = \frac{1}{d_{i,x}(\mathbf{a})}$ as the two functions have the same image on all possible subsets of the domain of the functions.

b) $\rightarrow$ *a)*
Suppose $r_{i,x}^{EP}(\mathbf{a}) = \frac{1}{d_{i,x}(\mathbf{a})}$ holds.
First, for $d_{i,x}(\mathbf{a}) = \infty$, $r_{i,x}^{EP}(\mathbf{a}) = \frac{1}{\infty} = 0$ and for $d_{i,x}(\mathbf{a}) = 1$, $r_{i,x}^{EP}(\mathbf{a}) = \frac{1}{1} = 1$. Thus, P.1

holds.
Second, $\frac{r^{EP}_{i,x}(\mathbf{a})}{r^{EP}_{j,x}(\mathbf{a})} = \frac{\frac{1}{d_{i,x}(\mathbf{a})}}{\frac{1}{d_{j,x}(\mathbf{a})}} = \frac{d_{j,x}(\mathbf{a})}{d_{i,x}(\mathbf{a})}$. Thus, P.2 holds. □

Proof of Proposition 1.
Part (1)
The first part of Proposition 1 provides the conditions for which there exists a unique subgame perfect Nash equilibrium in which all stage-1 agents take action $\underline{a}$. This is the case if any agent who takes action $\overline{a}$ always has an incentive to deviate to action $\underline{a}$, taking into account that agent K will allocate in stage 2 according to 3.8. An agent will deviate from $\overline{a}$ to $\underline{a}$, if

$$\pi_i^I(\overline{a}) + \frac{\rho}{c}\, r_{i,\overline{e}}(\overline{a}, a_{-i})\; j(\overline{e}) < \pi_i^I(\underline{a}) + \frac{\rho}{c}\, r_{i,\underline{e}}(\underline{a}, a_{-i})\; j(\underline{e})$$

Note that, in this setting, we can express the distance measure in terms of the number of agents that take $\underline{a}$ such that $d_{i,\underline{e}}(\underline{a}, a_{-i}) = 1 + |\sum_{i\in I} \mathbb{1}(a_i = \underline{a}) - t|$ and $d_{i,\overline{e}}(\overline{a}, a_{-i}) = 1 + |\sum_{i\in I} \mathbb{1}(a_i = \underline{a}) - (t-1)|$. Importantly, an agent who switches from playing $\overline{a}$ to playing $\underline{a}$ increases the sum of agents who play $\underline{a}$ by one. Therefore, for any given a_{-i}, $d_{i,\overline{e}}(\overline{a}, a_{-i}) = d_{i,\underline{e}}(\underline{a}, a_{-i})$ and $r_{i,\overline{e}}(\overline{a}, a_{-i}) = r_{i,\underline{e}}(\underline{a}, a_{-i})$.

Using this and solving for ρ yields

$$\rho < \frac{c}{r_{i,\overline{e}}(\overline{a}, a_{-i})} \cdot \frac{\pi_i^I(\underline{a}) - \pi_i^I(\overline{a})}{j(\overline{e}) - j(\underline{e})}$$

The RHS of the inequality is decreasing in agent i's causal responsibility for event $\overline{e}$. Therefore, when ρ is small enough such that an agent will deviate from $\overline{a}$ to $\underline{a}$ even when he is fully causally responsible for $\overline{e}$ ($r_{i,\overline{e}}(\overline{a}, a_{-i}) = 1$), then any agent will deviate from $\overline{a}$ to $\underline{a}$ for any level of causal responsibility and, hence, all agents playing $\underline{a}$ and agent K allocating according to 3.8 is the unique subgame perfect Nash equilibrium.

Part (2)
The second part of proposition 1 states that, if ρ is large enough such that deviating from $\overline{a}$ to $\underline{a}$ is deterred in the case of full causal responsibility, then there exists an integer $m \geq 1$ such that, if a number of agents in the interval $[t-m, t+m-2]$ takes action $\underline{a}$, then those who take action $\overline{a}$ have no incentive to deviate.

The lower bound is given by two conditions: First, if $(t-m)$ agents take action $\underline{a}$, then none of the $n-(t-m)$ agents who take action $\overline{a}$ has an incentive to deviate from $\overline{a}$ to $\underline{a}$ (equation 3.9). Second, if $(t-m-1)$ agents take action $\underline{a}$, then the agents who take action $\overline{a}$ have an incentive to deviate (equation 3.10).

$$\pi_i^I(\overline{a}) + \frac{\rho}{c}\, \frac{j(\overline{e})}{1 + |(t-m) - (t-1)|} \geq \pi_i^I(\underline{a}) + \frac{\rho}{c}\, \frac{j(\underline{e})}{1 + |(t-m+1) - t|} \tag{3.9}$$

and

$$\pi_i^I(\overline{a}) + \frac{\rho}{c}\,\frac{j(\overline{e})}{1+|(t-m-1)-(t-1)|} < \pi_i^I(\underline{a}) + \frac{\rho}{c}\,\frac{j(\underline{e})}{1+|((t-m)-t)|} \tag{3.10}$$

Simplified, m is the unique integer that lies in the intervall $(\frac{\rho}{c}\,\frac{j(\overline{e})-j(\underline{e})}{\pi_i^I(\underline{a})-\pi_i^I(\overline{a})}-1, \frac{\rho}{c}\,\frac{j(\overline{e})-j(\underline{e})}{\pi_i^I(\underline{a})-\pi_i^I(\overline{a})}]$, or

$$m = \{y \in \mathbb{N} \mid \frac{\rho}{c}\,\frac{j(\overline{e})-j(\underline{e})}{\pi_i^I(\underline{a})-\pi_i^I(\overline{a})} - 1 < y \leq \frac{\rho}{c}\,\frac{j(\overline{e})-j(\underline{e})}{\pi_i^I(\underline{a})-\pi_i^I(\overline{a})}\}$$

The interval uniquely identifies m, as one and only one integer lies in it. That $m \geq 1$ follows immediately from the condition that $\rho \geq c \cdot \frac{\pi_i^I(\underline{a})-\pi_i^I(\overline{a})}{j(\overline{e})-j(\underline{e})}$.

For the upper bound of the interval, the two conditions are that, if $t+m-2$ agents take action $\underline{a}$, none of those who take $\overline{a}$ has an incentive to deviate (condition 3.11), but if $t+m-1$ take $\underline{a}$, then there is such an incentive (condition 3.12). The reason for $t-m$ in the case of the lower bound and $t+m-2$ in the case of the upper bound is that both yield the same distance from pivotality for event $\overline{e}$, $(t-1)-(t-m) = m-1$ and $(t+m-2)-(t-1) = m-1$.

$$\pi_i^I(\overline{a}) + \frac{\rho}{c}\,\frac{j(\overline{e})}{1+|(t+m-2)-(t-1)|} \geq \pi_i^I(\underline{a}) + \frac{\rho}{c}\,\frac{j(\underline{e})}{1+|(t+m-1)-t|} \tag{3.11}$$

and

$$\pi_i^I(\overline{a}) + \frac{\rho}{c}\,\frac{j(\overline{e})}{1+|(t+m-1)-(t-1)|} < \pi_i^I(\underline{a}) + \frac{\rho}{c}\,\frac{j(\underline{e})}{1+|(t+m)-t|} \tag{3.12}$$

Simplified, m is again the unique integer that lies in the interval $(\frac{\rho}{c}\,\frac{j(\overline{e})-j(\underline{e})}{\pi_i^I(\underline{a})-\pi_i^I(\overline{a})}-1, \frac{\rho}{c}\,\frac{j(\overline{e})-j(\underline{e})}{\pi_i^I(\underline{a})-\pi_i^I(\overline{a})}]$. The upper and the lower bound coincide, if $m = 1$. In this case, punishment is just enough to deter the pivotal agents from switching from $\overline{a}$ to $\underline{a}$, but one step away from pivotality, taking $\underline{a}$ is more profitable.

Part (2), i)
If $t-m > 0$, conditions 3.9 and 3.10 provide the conditions for an equilibrium in which $t-m$ vote for $\underline{a}$ and $n-(t-m)$ vote for $\overline{a}$ and no agent has an incentive to deviate. This proves part 2) i) of Proposition 1.

Part (2), ii)
If $t-m \leq 0$, the distance from pivotality required to make taking $\underline{a}$ profitable is bigger than t. Therefore, even when 0 agents take $\underline{a}$, no agent has an incentive to deviate. Thus, all n agents taking $\overline{a}$ and agent K allocating according to 3.8 is an equilibrium.

Part (2), iii)

In contrast to the lower bound, the upper bound does not constitute an equilibrium if $m > 1$, i.e. if upper and lower bound do not coincide. In this case, if exactly $t + m - 2$ agents take action $\underline{a}$, then any agent who does so has an incentive to deviate to $\overline{a}$ (condition 3.11). On the other hand, if more than $t + m - 2$ agents take action $\underline{a}$, then any agent who takes $\overline{a}$ has an incentive to deviate to $\underline{a}$ (condition 3.12 holds for $t + m - 1$ and any higher number of agents who take $\underline{a}$). Therefore, if $t + m - 2 < n$, then there exists an equilibrium with all n agents taking $\underline{a}$ and agent K allocating according to 3.8.

□

Proof of Corollary 1.
The integer m is the unique integer provided by condition

$$m = \{y \in \mathbb{N} \mid \frac{\rho}{c} \frac{j(\overline{e}) - j(\underline{e})}{\pi_i^I(\underline{a}) - \pi_i^I(\overline{a})} - 1 < y \leq \frac{\rho}{c} \frac{j(\overline{e}) - j(\underline{e})}{\pi_i^I(\underline{a}) - \pi_i^I(\overline{a})}\}$$

By definition $\rho > 0$, $c > 0$, $j(\overline{e}) - j(\underline{e}) > 0$, and $\pi_i^I(\underline{a}) - \pi_i^I(\overline{a}) > 0$.

As the length of the interval is equal to 1 and thus constant and there is only one integer in the interval, it is sufficient to show that the upper bound of the interval is increasing/decreasing in a variable in order to show that m is increasing/decreasing in that variable. The following partial derivatives prove the four statements in Corollary 1.

i) $\frac{\partial \frac{\rho}{c} \frac{j(\overline{e})-j(\underline{e})}{\pi_i^I(\underline{a})-\pi_i^I(\overline{a})}}{\partial \rho} = \frac{1}{c} \frac{j(\overline{e})-j(\underline{e})}{\pi_i^I(\underline{a})-\pi_i^I(\overline{a})} > 0$

ii) $\frac{\partial \frac{\rho}{c} \frac{j(\overline{e})-j(\underline{e})}{\pi_i^I(\underline{a})-\pi_i^I(\overline{a})}}{\partial c} = -\frac{\rho}{c^2} \frac{j(\overline{e})-j(\underline{e})}{\pi_i^I(\underline{a})-\pi_i^I(\overline{a})} < 0$

iii) $\frac{\partial \frac{\rho}{c} \frac{j(\overline{e})-j(\underline{e})}{\pi_i^I(\underline{a})-\pi_i^I(\overline{a})}}{\partial\, j(\overline{e})-j(\underline{e})} = \frac{\rho}{c} \frac{1}{\pi_i^I(\underline{a})-\pi_i^I(\overline{a})} > 0$

iv) $\frac{\partial \frac{\rho}{c} \frac{j(\overline{e})-j(\underline{e})}{\pi_i^I(\underline{a})-\pi_i^I(\overline{a})}}{\partial\, \pi_i^I(\underline{a})-\pi_i^I(\overline{a})} = -\frac{\rho}{c} \frac{j(\overline{e})-j(\underline{e})}{(\pi_i^I(\underline{a})-\pi_i^I(\overline{a}))^2} < 0$

□

Proof of Proposition 2.
In order to prove Proposition 2, I use backward induction. The optimal behavior in stage 2 by agent K is characterized in equation 3.8. In stage 1, the agents rationally anticipate the behavior of agent K in stage 2 and act in order to maximize their payoff from the whole game. Note that, by assumption, $0 < t - 1 < n$.

First, whenever less than t agents take $\underline{a}$, any agent who does so has an incentive to deviate if

$$\pi_i^I(\overline{e}) + \frac{\rho}{c}\,\frac{j(\underline{e})}{1 + |\sum_{i\in I}\mathbb{1}(a_i = \underline{a}) - t|} < \pi_i^I(\overline{e}) + \frac{\rho}{c}\,\frac{j(\overline{e})}{1 + |\sum_{i\in I}(\mathbb{1}(a_i = \underline{a}) - 1) - (t-1)|}$$

where $\mathbb{1}(a_i = \underline{a})$ is equal to 1 if $a_i = \underline{a}$ and zero otherwise. The condition simplifies to

$$j(\underline{e}) < j(\overline{e})$$

which always holds by definition. Thus, if less than t agents take $\underline{a}$, any agent who takes $\underline{a}$ has an incentive to deviate and any agent who takes $\overline{a}$ has no incentive to deviate (by the same condition). Therefore, there always exists a SPNE in which all stage-1 agents take $\overline{a}$ and agent K allocates according to 3.8.

Second, whenever more than t agents take action $\underline{a}$, then those who do so have an incentive to deviate, if

$$\pi_i^I(\underline{e}) + \frac{\rho}{c}\,\frac{j(\underline{e})}{1 + |\sum_{i\in I}\mathbb{1}(a_i = \underline{a}) - t|} < \pi_i^I(\underline{e}) + \frac{\rho}{c}\,\frac{j(\overline{e})}{1 + |(\mathbb{1}(a_i = \underline{a}) - 1) - (t-1)|}$$

which again simplifies to

$$j(\underline{e}) < j(\overline{e})$$

which always holds by definition. Hence, there is no equilibrium in which more than t agents take $\underline{a}$.

Third, whenever exactly t agents take action $\underline{a}$, then those who do so have no incentive to deviate, if

$$\pi_i^I(\underline{e}) + \frac{\rho}{c}\,\frac{j(\underline{e})}{1} \geq \pi_i^I(\overline{e}) + \frac{\rho}{c}\,\frac{j(\overline{e})}{1}$$

which simplifies to

$$\rho \;\leq c \cdot \frac{\pi_i^I(\underline{e}) - \pi_i^I(\overline{e})}{j(\overline{e}) - j(\underline{e})}$$

At the same time, if exactly t agents take action $\underline{a}$, then those who do take $\overline{a}$ have no incentive to deviate, if

$$\pi_i^I(\underline{e}) + \frac{\rho}{c}\,\frac{j(\overline{e})}{1 + |(\mathbb{1}(a_i = \underline{a})) - (t-1)|} \geq \pi_i^I(\underline{e}) + \frac{\rho}{c}\,\frac{j(\underline{e})}{1 + |(\mathbb{1}(a_i = \underline{a}) + 1) - t|}$$

which again simplifies to $j(\underline{e}) \leq j(\overline{e})$ which is always given.

Hence, if $\rho \leq c \cdot \frac{\pi_i^I(\underline{e}) - \pi_i^I(\overline{e})}{j(\overline{e}) - j(\underline{e})}$, there exists an equilibrium in which t agents choose action $\underline{a}$, event $\underline{e}$ is implemented, and agent K allocates according to equation (3.8). This proves part (2) of Proposition 2.

On the other hand, if $\rho > c \cdot \frac{\pi_i^I(\underline{e}) - \pi_i^I(\overline{e})}{j(\overline{e}) - j(\underline{e})}$, then there exists a unique equilibrium in which all agents take action $\overline{a}$, event $\overline{e}$ is implemented, and agent K allocates according to equation 3.8. This proves part (1) of Proposition 2.

□

Chapter 4

The Spillover Effect of Institutions on Cooperative Norms, Preferences, and Beliefs[1]

4.1 Introduction

The success of any society is partly determined by the laws and norms that govern it. These laws and norms are often in place to overcome social dilemma situations in which the individual member's incentives are diametrically opposed to what is best for the society at large (common pool resources, public goods, etc.). Institutions help to back up laws and norms by enforcing them directly or by punishing individuals that don't comply with them.[2] One important aspect of the real world is that these institutions are limited in scope - they can seldom monitor all relevant areas and behaviors. The effectiveness of an institution for the implementation of a law or norm depends therefore crucially on whether the members of society comply with it, even when they are not monitored. Additionally, the introduction of such institutions can have unintended consequences and

[1]This paper must be cited as: Engl, F., A. Riedl and R. A. Weber (2015): "The Spillover Effect of Institutions on Cooperative Norms, Preferences, and Beliefs," Working Paper. We thank Pedro Dal Bo, Guillaume Fréchette, Bernd Irlenbusch and Martin Kocher for valuable discussions and conference participants at the 2015 Social Norms and Institutions conference in Ascona and the 2015 London Experimental Workshop (LES) for helpful comments.

[2]There exists a large experimental literature on the ability of sanctioning institutions to enforce norm compliance as well as on the endogenous uptake of sanctioning institutions (Faillo, Grieco, and Zarri, 2013; Fehr, Fischbacher, and Gächter, 2002; Fehr and Fischbacher, 2004; Gächter and Fehr, 2000; Galbiati and Vertova, 2008; Gürerk, Irlenbusch, and Rockenbach, 2006; Kosfeld, Okada, and Riedl, 2009; Kroll, Cherry, and Shogren, 2007; Markussen, Putterman, and Tyran, 2014; Masclet, Noussair, Tucker, and Villeval, 2003; Ostrom, Walker, and Gardner, 1992; Putterman, Tyran, and Kamei, 2011; Reuben and Riedl, 2013; Sutter, Haigner, and Kocher, 2010; Walker, Gardner, Herr, and Ostrom, 2000).

even backfire, leading to lower compliance in areas beyond its scope.[3] Understanding the mechanisms through which institutions foster or deter people's compliance with laws and norms when they are not monitored is therefore of great importance to public policy and economics.

For example, the government's ability to check the correctness of tax payments differs depending on the type of the tax. While income tax payments can be perfectly monitored in many countries, wealth taxes often rely on the voluntary provision of information. The institution that governs income taxes can therefore have positive or negative influence on the voluntary payment of wealth taxes. Relatedly, the police force cannot control all areas or record every transgression. Its behavior in those domains that can be monitored will affect how law-abiding citizens are, when they cannot be caught. In other areas, different but related behavior is monitored to a different extent. One such area is the state's aim to promote energy conservation. While the state can implement laws that make it obligatory to comply with energy conservation in some areas, e.g. by phasing out the supply of traditional light bulbs[4] or by setting mission reduction targets for new cars[5], these policies only reach their goal if they change people's behavior per se, i.e. if people internalize energy conservation as a norm to be followed, e.g. by reducing the usage of cars in favor of public transport, even when there is no direct obligation to do so.

In this paper, we study in a laboratory experiment whether social norms that are monitored and enforced by an institution in one domain spill over to domains without such an institution. Additionally, we explore whether the way in which an institution is formed matters for the extent and direction of the norm compliance spillover. We are particularly interested whether institutions that are endogenously adopted through a democratic process facilitate the spillover compared to institutions that are exogenously imposed by an external authority. Finally, we study if and how the presence of an institution directly alters pro-social preferences and beliefs about others' cooperativeness and whether potential effects extend to new counterparts.

To this end, we designed a laboratory experiment in which subjects simultaneously play two standard public goods games, within identical groups, repeatedly for 20 periods. In a "no institution" treatment both games are non-monitored and subjects can freely choose their contribution. In the other two treatments, one of the two games is monitored by an institution ("PGG Right"), the other is not ("PGG Left"). The institution that

[3]Such unintended consequences have been shown for environments that are actively regulated by an institution. For example, Falk and Kosfeld (2006) demonstrate that introducing incentives which are supposed to increase a worker's effort, actually lead to a decrease because they undermine the worker's feeling of being trusted. Frey (1993) argues that crowding out of intrinsic motivation might spill over to areas in which such incentives are not present.

[4]`http://ec.europa.eu/energy/lumen/index_en.htm`

[5]`http://ec.europa.eu/clima/policies/transport/vehicles/cars/index_en.htm`

we consider is a central authority that punishes subjects that contribute less than a specified amount (*minimum contribution requirement*, henceforth *MCR*). The penalty is the loss of all income in that period from the respective game. Depending on the treatment the institution is either endogenously adopted or exogenously imposed. In the "exogenous institution" treatment, the *MCR* is set to the full endowment. In the "endogenous institution" treatment, the *MCR* is determined by a voting mechanism. Each subject submits a vote for his preferred *MCR*. The implemented *MCR* in a group is the minimal number that any member voted for. By comparing the contributions in the non-monitored public goods game across the three treatments, we can observe i) whether there exists a spillover of institution-backed norms to non-monitored environments and ii) whether the procedure with which an institution is implemented affects the spillover. Furthermore, our design includes additional parts that measure the lasting impact of institutions on beliefs about others' cooperativeness and pro-social preferences. Both beliefs and preferences are elicited with respect to familiar as well as unfamiliar group members. A number of additional elicited individual characteristics allow us to disentangle several potential channels.

We find that the presence of an institution that regulates cooperative behavior in one domain generally enhances cooperation in domains that lie beyond the scope of the institution compared to a situation in which no institution was present in either domain. As a result, institutions that regulate cooperation induce significantly positive indirect welfare effects in addition to their immediate welfare effect in the regulated environment. However, we also find that the institution formation procedure matters for the stability of the spillover. Cooperation boosted by an exogenously imposed institution nevertheless decays over time, while the endogenously adopted institution leads to stable voluntary cooperation levels in domains beyond its scope. Additionally, we find that institutions have positive effects on cooperative behavior, preferences for cooperation and beliefs about others' cooperativeness that persist even after they cease to exist, compared to environments in which an institution never existed. These effects extend to new counterparts. Overall, we therefore conclude that institutions which regulate cooperative behavior have, in our study, unambiguously positive effects on voluntary cooperative behavior in domains that lie beyond the immediate scope of the institution and that endogenously implemented institution have an advantage over exogenously imposed institution with regard to the stability of the spillover effects. When deciding about the implementation of such institutions in reality, these effects should to be taken into account and can, potentially, alter the analysis in favor of (endogenously implemented) institutions.

The paper is organized as follows: Section 2 summarizes the related literature and develops hypotheses regarding the effect of institutions on behavior outside of their scope. Section 3 introduces the experimental design. Results are presented in Section 4. In Section 5,

we discuss possible explanations for our findings in light of the hypotheses of Section 2 and Section 6 concludes and discusses directions for future research.

4.2 Related literature and hypotheses

Our study is related to a literature in experimental economics that examines behavior in simultaneously and sequentially played games. For example, studies have shown that strategies that subjects develop while playing simple games are also applied to more complex, simultaneously played games (Bednar, Chen, Liu, and Page, 2012), that participation in a public goods game reduces overbidding in simultaneously played lottery contests (Savikhin and Sheremeta, 2013), that pro-social behavior is greater if two simultaneously played public goods games are played with different groups as when they are played with identical groups (McCarter, Samak, and Sheremeta, 2013, WP), and that there is only a small behavioral spillover between two public goods games that are played with different group members (Falk, Fischbacher, and Gächter, 2013). Blackwell and McKee (2003) also study simultaneously played public good games where one local group is entailed in a larger global group and with varying MPCRs between groups. They find that increasing the MPCR in the global group increases contributions at the expense of the private account, not at the expense of contributions to the local group. In a similar setting, Fellner and Lünser (2014) show that higher returns in global groups are not sufficient to guarantee cooperation when local groups provide an information advantage about individual contributions. Bernasconi, Corazzini, Kube, and Maréchal (2009) find that splitting one public goods game into two simultaneously played public goods games increases subject's contributions.

In the case of sequentially played games, studies show that groups that manage to sustain high efficiency levels in a weak-link game have higher cooperation rates in subsequently played prisoner's dilemma games (Knez and Camerer, 2000), that there are learning spillovers between strategically similar games (Grimm and Mengel, 2012), that there exist only very modest spillovers between competitive and cooperative games that are played with the same opponents (Cason and Gangadharan, 2013). Furthermore, Cason, Savikhin, and Sheremeta (2012) find behavioral spillovers between minimum- and median-effort coordination games when they are played sequentially, but not when they are played simultaneously. Herz and Taubinsky (2013, WP) find that earlier experience in ultimatum games with proposer or responder competition influences subsequent minimum acceptance offers of responders in standard ultimatum games. Falk, Fehr, and Zehnder (2006) show the influence of minimum wages on subsequent reservation wages.

Our study differs from the aforementioned ones as it focuses on the effect that an institution, which regulates cooperative behavior in one environment, has on behavior in another environment, without such an institution. Prior experimental research in economics has demonstrated the effectiveness of institutions in enforcing high cooperation levels in social dilemma situations (Ostrom, Walker, and Gardner, 1992; Gächter and Fehr, 2000) and studied the endogenous uptake of such institutions by society (Gürerk, Irlenbusch, and Rockenbach, 2006; Kosfeld, Okada, and Riedl, 2009). These institutions typically work by changing the monetary incentives for non-cooperative behavior, making it more costly relative to cooperative behavior. However, as argued in the introduction, in the real world, such institutions are often limited in scope. If institutions only influence relative prices in one domain, then they should not influence behavior in environments beyond their scope, where relative prices are unchanged. Furthermore, if an institution only changes relative prices, changes in behavior should not persist once the institution ceases to exist. This yields our null hypothesis:

Hypothesis 1 (Null)**.** *The presence of an institution in one domain does not affect cooperative behavior in domains beyond the scope of the institution.*

This null hypothesis is supported under the standard preference framework of pure selfishness as well as under the assumption of outcome-based social preferences. Applied to our experimental framework, if players are purely selfish, they should contribute zero in "PGG Left" independent of the presence or the type of the institution that governs "PGG Right" as contributing zero maximizes their own monetary payoff.[6] Similarly, the presence of an equilibrium with positive contributions in "PGG Left" according to outcome-based social preferences in the formulation of Fehr and Schmidt (1999) does not depend on the presence or the type of an institution in "PGG Right".[7]

However, if institutions have an impact beyond changing monetary incentives, it is possible that cooperative behavior in environments outside the scope of the institution is influenced by its presence and type. The null hypothesis thus serves as a benchmark against which to test such non-standard effects of institutions. In the following, we develop hypotheses regarding the potential impact of an institution on cooperative behavior beyond its scope through its effect on people's preferences and beliefs.

Several studies have argued that institutions can affect behavior, in addition to their influence on monetary incentives, through a crowding out or crowding in of intrinsic incentives (for survey articles, see Frey and Jegen (2001), Gneezy, Meier, and Rey-Biel

[6]For a detailed summary of the theoretical predictions of standard preferences, see Appendix 4.A.1

[7]We refer the reader to Proposition 4 in the paper of Fehr and Schmidt (1999). Importantly, the conditions on the parameters for which any positive contribution level can be sustained as an equilibrium are not affected by our setup with two separate contribution decisions, or the implemented *MCR* in "PGG Right". If cooperation can exist in equilibrium, the theory is silent about the level of cooperation.

(2011), and Bowles and Polanía-Reyes (2012)).[8] If institutions affect intrinsic preferences, then this change of preferences will spill over and influence a person's behavior even in domains in which the institution is not active (Frey, 1993). In the following, we discuss several reasons why that might be the case for the kind of institutions that are active in our experiment.

On the one hand, institutions can decrease people's intrinsic willingness to act prosocially. For example, people might be averse against exogenous control of their behavior because it compromises their sense of authority (Deci, 1975; Deci and Ryan, 1985; Deci, Koestner, and Ryan, 1999). Hence, an institution that enforces prosocial behavior might lead people, even those who are intrinsically motivated to act prosocially, to resist the rule when possible. While such resistance is not possible in domains that are regulated by the institution, it can manifest itself in domains beyond the scope of the institution. In our setting, such strict exogenous control of behavior is only active in the "PGG Right" of the "exogenous institution" treatment, in which contribution of the full endowment is exogenously enforced by a harsh punishment rule. Therefore, such resistance to exogenous control can show up in lower contribution levels in the "PGG Left" compared to the setting without an institution or when the institution was endogenously adopted.[9]

On the other hand, institutions can also increase people's intrinsic willingness to act prosocially. For example, Krupka and Weber (2013) hypothesize that people have a preferences to follow known social rules or norms.[10] If the *MCR* signals such a social

[8]Previous studies mostly compare behavior in the absence of an institution with behavior in the presence of an institution. For example, Bowles and Polanía-Reyes (2012) interpret results of Irlenbusch and Ruchala (2008) as evidence of crowding out because, after the introduction of a monetary incentive, prosocial behavior should have risen even more than it did, compared to the absence of monetary incentives. Our experimental design provides a cleaner test of such an effect because we directly elicit intrinsic preferences for cooperation both before and after an institution was active. To the best of our knowledge, our study is first that provides such clean evidence on the causal effect of institutions on intrinsic preferences for cooperation.

[9]It has also been argued that institutions can signal distrust and that the feeling of being distrusted crowds out intrinsic motivation (Frey, 1997; Gneezy, Meier, and Rey-Biel, 2011). For example, Falk and Kosfeld (2006) provide evidence for such a channel in an experiment in which a principal could either let an agent freely choose an effort level or enforce a minimum level of effort. They find that agents provide lower effort when a minimum level was enforced than when they were free to choose and that subjects felt mistrusted when a minimum effort was enforced (see also Tausch (2014)). Belot and Schröder (2015) provide field experimental evidence that such an effect might spill over to other domains. They find that monitoring workers in one dimension of their task (work quality) has detrimental effects on their performance in another, non-monitored dimensions of their task (punctuality) compared to when there was no monitoring of the first dimension. In our setting, subjects might interpret votes for a high *MCR* in the "endogenous institution" treatment as a signal of distrust, and react by lowering their contributions in "PGG Left". However, since institutions are implemented by mutual agreement and the induced higher cooperation benefits all group members and not only a principal, we don't expect this channel to play a major role.

[10]Relatedly, the literature on the expressive function of laws proposes that laws can change behavior not only through changing incentive (the sanction function of law), but also through changing the norms that people obey (the expressive function of law) (Cooter, 1998; Funk, 2007; Kahan, 1998; Sunstein, 1996).

rule, then subjects might prefer to comply with it, even in the unregulated "PGG Left". Such an effect would supposedly be stronger in the "endogenous institution" treatment when the *MCR* is implemented by the voting process and thus reflects a social rule that everyone in the group agreed to.

Relatedly, Rand, Peysakhovich, Kraft-Todd, Newman, Wurzbacher, Nowak, and Greene (2014) propose a theory of human cooperation which argues that cooperativeness is driven by intuitive heuristics, whereas own payoff maximizing behavior is driven by deliberative thinking (Social Heuristics Hypothesis, henceforth SHH). Peysakhovich and Rand (forthcoming) demonstrate how such intuitive heuristics can be shaped by the environments. In stage 1 of their experiments, subjects play an infinitely repeated game that establishes, depending on the treatment, either a cooperative or a non-cooperative norm. In stage 2, subjects play a range of punishment games. They find that subjects who established a cooperative norm in stage 1 are more likely to punish selfishness in stage 2. They also show that subjects who score lower on the CRT, and are thus more likely to rely on intuitive responses, are especially prone to exhibit the effect, which is interpreted as support of the SHH. In our setting, the *MCR* in the "exogenous institution" treatment and the "endogenous institution" treatment provides a simple heuristic of cooperativeness that is easily applied: contribute as much as in "PGG Left" as the *MCR* dictates for "PGG Right".

Together, these arguments suggest that institutions can have an effect on preferences for cooperation and that the effect might differ depending on whether the institution is exogenously imposed or endogenously adopted. However, since the direction of the effect is not unambiguous, our hypothesis posits an effect, but leaves open the direction.

Hypothesis 2 (Preference effect)**.** *The presence of an institution that enforces cooperation increases or decreases intrinsic preferences for cooperation.*

A second channel through which an institution can affect cooperative behavior in the unregulated "PGG Left" is its effect on beliefs about others' cooperativeness. It has been established that many people act as conditional cooperators in public goods games (Fischbacher, Gächter, and Fehr, 2001; Fehr and Fischbacher, 2004). That is, they are willing to cooperate if their group members also cooperate, but unwilling to do so if their group members don't cooperate. If this is the case, beliefs are an important determinant of cooperative behavior as the higher a person's beliefs about his group members' cooperativeness the more he will cooperate himself (Fischbacher and Gächter, 2010). Therefore, if institutions directly change beliefs about others' cooperativeness, this would provide a channel through which cooperative behavior is affected even beyond the scope of the

institution.[11]

In the "endogenous institution" treatment, beliefs about others' cooperativeness in "PGG Left" can be influenced by the *MCR* if the voting mechanism provides subjects with an opportunity to learn about their group members' characteristics. This could be the case if subjects interpret, e.g., a fellow group member's vote for an *MCR* of 20 as a signal that that group member understands the game and agrees that high contribution levels are what the group should strive for. If all group members vote for 20, the whole group signals understanding of the game and the socially efficient solution. Such knowledge about the group members' understanding of what should be played can increase beliefs about their cooperativeness in "PGG Left" in which a similar problem has to be solved. If an increase in beliefs translates into higher cooperativeness, the effect could be interpreted as an institution's effect on trust. Indeed, it has also been argued that institutions can have a positive effect on trust and trustworthiness (Tabellini, 2008). For example, Cassar, D'Adda, and Grosjean (2014) finds experimental evidence that the presence of an exogenously implemented institution that increases honest behavior increases trust levels even after the institution stopped existing.

On the other hand, theoretical work in the principal-agent literature suggests that endogenously implemented institutions can provide a signal for the type-distribution of agents that goes in the other direction (Sliwka, 2007; van der Weele, 2009; Benabou and Tirole, 2011, WP). Specifically, an institution can signal a high proportion of selfish agents if the person implementing the institution has knowledge of the type-distribution. Galbiati, Schlag, and van der Weele (2013) test this hypothesis in a lab experiment in which sanctions could either be implemented exogenously, or by a subject with information advantage about the other subjects' cooperativeness. They find that cooperative subjects perceive actively chosen sanctions as a negative signal about the types of the other subjects to which they react by lowering their cooperativeness. In our setup, if a subject votes for a high *MCR*, this can signal that he believes that his group members are selfish. Thus, his vote can influence the beliefs of the rest of the group either because they trust this judgment better than their own and update their own beliefs, or because they know that, if the other player is conditionally cooperative and thinks he plays with selfish group members, he will cooperate little in "PGG Left".

The game-theoretical literature on play in separate games suggests that, when playing sufficiently complex games, players categorize games into equivalence classes. Within one equivalence class, players then reduce complexity by bundling some aspect across all

[11]Note that the institutionally regulated contributions in "PGG Right" themselves don't lead to increased contributions for conditional cooperators in "PGG Left" as the conditional cooperators themselves are also forced to contribute the same amount as their group members. Only if their conditional cooperation schedule would have a slope above one, which is empirically seldom observed, would they increase their contribution in "PGG Left".

games. For example, players can bundle their beliefs about the behavior of the other players in an equivalence class (Jehiel, 2005). Evidence in support of this hypothesis was found in Huck, Jehiel, and Rutter (2011). In our setting, if players think of the two simultaneously played public goods games as belonging to one equivalence class, belief bundling implies that players believe that the average contribution of the other players is the same in the two simultaneously played games. Having an institution in the regulated game can thus raise the belief about average contributions in the unregulated game.[12]

Together, these arguments suggest that institutions, especially endogenously implemented ones, can have an effect on beliefs about others' cooperativeness. However, since the direction of the effect is not unambiguous, our hypothesis again posits an effect, but leaves open the direction.

Hypothesis 3 (Belief effect)**.** *The presence of an institution that enforces cooperation in one domain increases or decreases beliefs about the cooperativeness of group members in domains beyond the scope of the institution.*

Both, the preference and the belief channel can lead to an effect of an institution in "PGG Right" on cooperative behavior in "PGG Left".[13] In addition, there exist channels that predict a change of behavior in "PGG Left" independent of a change in preferences or beliefs. For example, pure altruism theory suggests that people want a certain amount of public good provided and are indifferent whether it is provided through their own contribution or those of others. Therefore, the theory would suggest that an institution, which enforces contributions in the "PGG Right", crowds out voluntary contributions to "PGG Left" one-to-one if the institutionally enforced level lies above what the subject wanted to achieve (Bernheim, 1986; Andreoni, 1988). Similarly, if people have preferences of impure altruism and care about their own contribution to the public good, the institution in "PGG Right" will lead to incomplete crowding out of own contributions (Andreoni, 1990). Together, these effects give rise to the main hypothesis on the behavioral effect of institutions. In Section 5, we will return to discuss our results in light of the different possible channels.

Hypothesis 4 (Behavioral effect)**.** *The presence of an institution that enforces cooperation in one domain increases or decreases cooperative behavior in domains beyond the scope of the institution.*

Some existing evidence from sequentially played games suggests that there is, indeed, a positive effect of monetary incentive for cooperativeness on cooperation even after the

[12]For other types of bundling, see Grimm and Mengel (2012).

[13]Note that there also exist feedback mechanisms between beliefs and preferences. If a group member thinks the others' preferences have changed, this will also change his belief about their behavior. Similarly, if a group member's belief about others' cooperativeness increases, this might change his preferences if he wants to conform with the group.

monetary incentive is removed. For example, Brandts and Cooper (2006) find that while increasing the monetary incentives for cooperation in repeatedly played weak-link games increases cooperative behavior, a decrease in monetary incentives has only little effect on subsequent cooperation rates, thus leading to higher cooperation rates compared to a situation in which the monetary incentive was never present. In a similar vein, Galbiati and Vertova (2008) study how altering a minimal contribution requirement affects contributions in repeated public goods games. They, too, find that an increase in minimum contribution requirements leads to increased contributions but that a decrease does not lead to a similar decrease in contributions.

A series of studies has shown that endogenously adopted rules are more effective in instilling socially desirable behavior in social-dilemma situations than exogenously imposed rules (Dal Bó, Foster, and Putterman, 2010; Sutter, Haigner, and Kocher, 2010; Tyran and Feld, 2006). Most related to our study, Kamei (2014, WP) studies two simultaneously played public goods games with endogenously or exogenously implemented institutions. The endogenously implemented institution was either present in both games, or one of the two games was monitored by the endogenous and the other by the exogenous institution. He finds that, if an institution was endogenously implemented in one game, this had positive effects on contributions in both games, thereby demonstrating a spillover effect between two institutionally governed environments.

4.3 Experimental Design

The experiment consists of five parts. The main part (Part II) measures the extent to which a norm for cooperation that is monitored and enforced by an institution in one environment spills over to an identical environment that is not monitored by such an institution. The treatments in Part II vary the way in which the institution is implemented in order to distinguish between spillovers that are generated by endogenously adopted versus exogenously imposed institutions. Parts I, III and IV are preference and belief elicitation stages that help to disentangle whether treatment effects are driven by changes in subjects' beliefs or preferences. These parts are identical across all treatments. Finally, in Part V, individual characteristics are collected. An overview of the experimental design is provided in Table 4.1.

The social dilemma situation considered is a standard public goods game. Across all parts and treatments, the following characteristics of the public goods game are held constant. It is played in groups of four members ($n = 4$). Each group member is endowed with 20 points ($w = 20$) and can decide how many points to keep for himself and how many to contribute to a public good. The sum of points contributed to the public good is doubled

and equally distributed among all members of the group ($a = 0.5$). We chose a "relatively high" MPCR of 0.5 in order to incite some level of cooperation also in the baseline which enables us to also detect a potential crowding out effect on cooperation. Thus, given the contribution of all group members ($\mathbf{g} = (g_1, ..., g_4)$) the material payoff of group member i in the standard game is equal to

$$\pi_i(\mathbf{g}) = 20 - g_i + 0.5 \sum_{j=1}^{4} g_j$$

Before explaining Part I, III, IV, and V in more detail, we begin with describing Part II, which is the main part of the experiment.

Table 4.1: Overview of experimental design

Part I	Preference and belief elicitation *(randomly determined group)*		
Part II	20 periods of 'PGG Left' and 'PGG Right' *(new group - absolute stranger matching)*		
	No institution	**Exogenous institution**	**Endogenous institution**
Part III	Preference and belief elicitation *(same group as in Part II)*		
Part IV	Preference and belief elicitation *(new group - absolute stranger matching)*		
Part V	Individual characteristics		

4.3.1 Part II: Treatment stage

At the beginning of Part II, subjects were randomly matched into groups of four subjects with whom they had not interacted with before (absolute stranger matching). Within Part II, subjects played repeatedly, for 20 periods, with the same group of subjects (partner matching). Part II differs between three treatments which are called "no institution", "exogenous institution", and "endogenous institution" treatment.

"No institution" treatment

In each period of the "no institution" treatment, subjects simultaneously played two public goods games with the same group members. The parameters of both games are as specified before. The two public goods games were displayed next to each other on the

same computer screen. In the following, these games are called "PGG Left" and "PGG Right".[14]

Before subjects made their contribution decision in the two public goods games, they were asked to indicate, separately and in each period, their belief about the average contribution of the other three group members in the two games. Belief elicitation was not incentivized monetarily, but subjects were asked to enter their best estimates.

After that, subjects indicated, separately, their contribution to the public good in both games. Subjects were endowed with 20 points for each game and were free to contribute any integer value between zero and 20 points to each game.

At the end of each period, subjects were informed about the contributions of all group members to the public good and their payoffs from both games. Contributions were displayed in descending order and it was not possible to identify which member of the group contributed which number of points to the public good in the two games. The payoff of each period consisted of the sum of the payoffs of the two games. At the end of the experiment, the payoff of one of the 20 periods was randomly selected to be paid out to the subjects. Specifically, the payoff for the randomly selected period was multiplied by 20, so that it counted for all 20 periods.

"Exogenous institution" treatment

We implemented two treatments in order to test our hypothesis. In both treatments, the setup of "PGG Left" is identical to the "no institution" treatment, i.e. subjects were free to contribute any integer amount between zero and 20 points to the public good. The payoff structure of "PGG Right", however, is affected by the treatments. In particular, "PGG Right" is governed by an institution that monitors the group members' contributions in "PGG Right" and punishes those members that contribute less than a certain *minimum contribution requirement* (henceforth *MCR*). Specifically, the income from "PGG Right" of any group member who contributes at least as many points to the group account as specified by the *MCR* is not affected by the *MCR*, but any group member who contributes fewer points to the group account than the minimum level specified by the *MCR* loses any income from "PGG Right" in that period. In the "exogenous institution" treatment, the *MCR* is set to 20. The payoff from "PGG Right" in the "exogenous

[14] In the experiment, the two games were neutrally labeled as "Task Left" and "Task Right".

institution" treatment treatment is thus determined as follows:

$$\pi_i(\mathbf{g}) = \begin{cases} 20 - g_i + 0.5 \cdot \sum_{j=1}^{4} g_j & \text{if } g_i = 20 \\ 0 & \text{if } g_i < 20 \end{cases} \tag{4.1}$$

Note that, if one group member is penalized for contributing less than the contribution threshold in "PGG Right", the incomes of the other group members are not affected. Thus, the other group members still benefit from any contributions made by any group member in "PGG Right".

Again, the total per-period payoff of each subject was equal to the sum of the payoffs in "PGG Left" and "PGG Right". Also, the payoff of one the 20 periods was randomly selected to be implemented and multiplied by 20.

"Endogenous institution" treatment

The "endogenous institution" treatment consists of two stages that are repeated in every period: an *institution formation stage* and a *contribution stage*.

Institution formation stage: The institution is identical to the *MCR* in the "exogenous institution" treatment. The only difference is that, in the "endogenous institution" treatment, each group endogenously selects the *MCR* for that group in each period, instead of facing an exogenously set contribution threshold of 20. Each member of the group casts a vote on which *MCR* it would like to be implemented by specifying an integer value between zero and 20. After all votes are collected, the lowest contribution threshold that was voted for by any group member is implemented as the *MCR* for that period. This mechanism ensured unanimity in the sense that each group member agreed that the *MCR* should be at least as high as the implemented one.

At the end of the *institution formation stage*, the subjects were informed about the implemented *MCR* for that period and the votes that were cast. Votes were displayed in descending order and it was not possible to identify which member of the group voted for which *MCR*.

After the subjects were informed about the *MCR* but before they made their contribution decisions, they were asked to indicate their belief about the other group members' average contribution to the two public goods games.

Contribution stage: The contribution stage is identical to the "exogenous institution" treatment with the only difference that the *MCR* of each period is endogenously selected

in the *institution formation stage* and not exogenously set to 20. Hence, the payoff from "PGG Right" in the "endogenous institution" treatment is thus determined as follows:

$$\pi_i(g_1, ..., g_4) = \begin{cases} 20 - g_i + 0.5 \cdot \sum_{j=1}^{4} g_j & \text{if } g_i \geq MCR \\ 0 & \text{if } g_i < MCR \end{cases} \tag{4.2}$$

Again, the total per-period payoff of each subject was equal to the sum of the payoffs in "PGG Left" and "PGG Right". Also, the payoff of one the 20 periods was randomly selected to be implemented and multiplied by 20.

4.3.2 Part I, III & IV: Preference and belief elicitation stages

We elicited cooperative preferences and beliefs about others' cooperativeness (i) before the main task (Part I), (ii) afterward with respect to the identical group from Part II (Part III) and (iii) with respect to a new group of randomly-selected participants that they had never interacted with before (Part IV). In these parts, there is no institution and the parts are identical across all three treatments. The elicitation tests whether institutions influence preference and beliefs even after they cease to exist and even towards new counterparts that subjects had never interacted with before. Furthermore, it helps to disentangle whether a spillover is driven by changes in subjects' preferences or by changes in subjects' beliefs about the cooperativeness of others.

Each part consisted of two stages: a *belief-elicitation stage* and a *preference-elicitation stage*. In the *belief-elicitation stage*, we elicited subjects' beliefs about the contribution decision of the other three group members in the standard public goods game with the Truncated Interval Scoring Rule introduced by Schlag and van der Weele (2012). Specifically, subjects were asked to provide two integer values as the upper and the lower bound of the range of values that they believe will contain the actual average contribution of the other group members (rounded to the nearest integer). Subjects could earn 20 ECU if they specified a range that consists of only one number and that number was equal to the actual rounded average unconditional contribution of the other group members in the preference-elicitation stage. For each unit that the provided range increased in width, a subject's potential earnings decreased by one ECU. Subjects earned nothing if the actual average contribution of others lied outside of the range they specified. Thus, subjects were incentivized to reveal their true beliefs as precisely as possible. The width of the range they provided is a measure of how certain they were of the correctness of their beliefs.

While unconditional contributions are informative about cooperative behavior, they are

not informative about a subject's preference for cooperation. In the *preference-elicitation stage*, we therefore employ a variant of the strategy method as introduced by Fischbacher, Gächter, and Fehr (2001) to elicit subjects' conditional contributions. This procedure has the advantage that, in contrast to observations of unconditional contribution decisions, conditional contribution decisions are not confounded with beliefs, and are thus informative about a subject's true preference for cooperation. Subjects first indicated their unconditional contribution to a standard public good game. Then, subjects were asked to specify how much they would contribute *for each of the 21 possible levels of average contribution* (rounded to integers) of the other group members. Three of the four members of a group were then randomly selected to implement their specified unconditional contribution; for the last member of the group his conditional contribution decision was implemented given the rounded average of the other group members' unconditional contributions.

4.3.3 Part V: Individual characteristics

In Part V, we collected individual characteristics from each of the participants. First, we elicited cognitive ability by use of the Cognitive Reflection Test (Frederick, 2005) and rule-following propensity via the rule-following task introduced in Kimbrough, Miller, and Vostroknutov (2014). In the rule-following task, subjects saw, on their computer screen, two baskets, one yellow and one blue, and a ball. They were told that they would earn 2 ECU if they place the ball in the yellow basket and 1 ECU if they place the ball in the blue basket. However, they were also told that the rule is to place the ball in the blue basket. This procedure was repeated for 30 balls. The number of balls placed in the blue basket is informative about a subject's propensity to follow an arbitrary rule that does not have any payoff consequences.

Additionally, we asked the subjects a series of questions in order to elicit their attitudes towards risk, patience, altruism, reciprocity and trust. These questions were English translations of the ones included in several waves of the German Socio Economic Panel (SOEP) survey.[15] The behavioral validity of the risk and patience question was established with incentivized experiments in Dohmen, Falk, Huffman, Sunde, Schupp, and Wagner (2011) and Vischer, Dohmen, Falk, Huffman, Schupp, Sunde, and Wagner (2013).[16]

At the very end of the experiment, we asked subjects about their age, gender, and academic major. At that point we also asked them about their reasoning when making the contribution decision for "PGG Left" and, in the "endogenous institution" treatment, about their reasoning when making the voting decision for the *MCR* in "PGG Left".

[15] All questions are reproduced in full in Appendix 4.A.4.

[16] Another example of the use of these questions is Becker, Deckers, Dohmen, Falk, and Kosse (2012).

4.3.4 General procedures

Before subjects entered the lab, they randomly drew a place card that specified at which computer terminal to sit. Subjects found paper copies of the consent form and the instructions for Part I at their assigned computer terminals. Subjects were informed that the experiment consists of several parts, but were not informed about the content of each part. At beginning of each part, the instructions of that part were read aloud to ensure common information regarding the content of the instructions. The instructions to Part I and Part II included comprehension questions that had to be answered correctly before the respective part could begin. The original English instructions for Part I and the "endogenous institution" treatment of Part II can be found in Appendix B, together with screen shots of the decision-relevant stages. All sessions were conducted in English.

We conducted six sessions on three consecutive days in November 2014 in Maastricht, Netherlands, with 136 subjects in total and six sessions on three consecutive days in February 2015 in Zurich, Switzerland, with 136 subjects in total. Overall, 272 subjects participated. Treatments were randomized across sessions and each treatment was run four times, twice in the morning and twice in the afternoon, twice in Maastricht and twice in Zurich. Each subject participated only once. Overall, 92 subjects participated in the "no institution" treatment, 88 subjects in the "exogenous institution" treatment, and 92 subjects in the "endogenous institution" treatment.

The sessions in Zurich took place at the Laboratory for Behavioral and Experimental Economics of the Department of Economics at the University of Zurich and the sessions in Maastricht took place at the Behavioral and Experimental Economics Laboratory (BEElab) of the School of Business and Economics at Maastricht University. The experiments were run with the software "z-Tree" (Fischbacher, 2007). We used the softwares "hroot" (Bock, Baetge, and Nicklisch, 2014) and "ORSEE" (Greiner, 2003) for recruitment. Subjects were students from the University of Zurich, the Swiss Federal Institute of Technology in Zurich and Maastricht University.

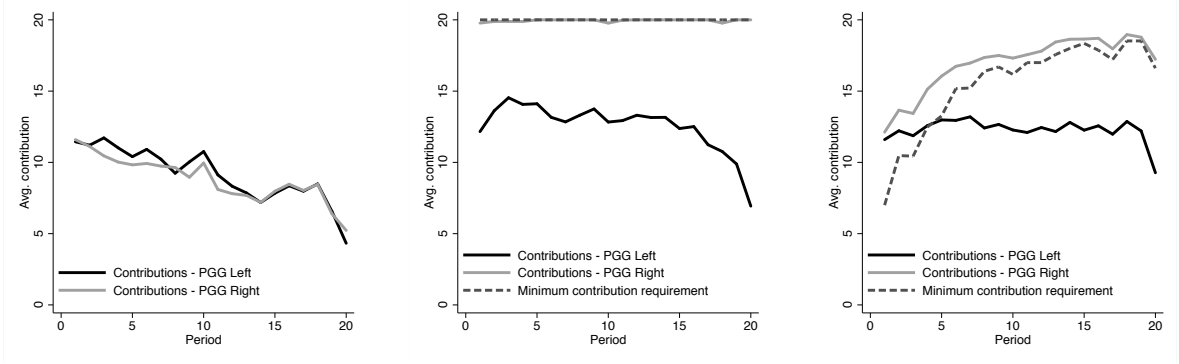
Sessions lasted about 2.5 hours. Payoffs from the experiment, denominated in "ECU," were converted into money at the rate of 65 ECU to €1 (about $1.25 at the time of the experiment) in Maastricht and 100 ECU to CHF 3 (about $3.25 at the time of the experiment) in Zurich. Subjects were paid out anonymously at the end of the experiment. On average, subjects earned €22.52 in Maastricht, with no show-up fee, and CHF 55.45 in Zurich , which included a show-up fee of CHF 10. The total payoff from the experiment equaled the sum of the payoffs in the five parts (plus the payment of a show-up fee in Zurich). Subjects learned about the results of and their payoffs from the five individual parts only at the very end of the experiment, after all decisions were made.

4.4 Results

This section presents the results of the experiment. First, Part II, the main part of the experiment, is analyzed and then Part I, III, and IV are analyzed.

4.4.1 Part II

Figure 4.1 gives an overview of behavior in the three treatments. For each treatment, it shows the average contributions to "PGG Left", "PGG Right", and the *MCR*. As can be seen, institutions are indeed effective in enforcing cooperative behavior in "PGG Right". Averaged over all periods, contributions to "PGG Right" are significantly higher in the "exogenous institution" treatment (19.95) and the "endogenous institution" treatment (16.95) compared to the "no institution" treatment (8.83) (Wilcoxon ranksum tests, p=0.000 and p=0.000).[17] The difference between the "exogenous institution" treatment and the "endogenous institution" treatment is also significant (Wilcoxon ranksum test, p=0.000). The correlation between the *MCR* and contributions to "PGG Right" is positive and significant (Spearman's rho, ρ=0.933, p=0.000). The *MCR* in "PGG Right" was violated in only 2 out of 1840 observations in the "endogenous institution" treatment and in only 7 out of 1769 observations in the "exogenous institution" treatment. Hence, subjects clearly understood the incentive effects of the *MCR* and a higher *MCR* led to higher contributions in "PGG Right".



(a) No institution (b) Exogenous institution (c) Endogenous institution

Figure 4.1: Contributions to "PGG Left", "PGG Right", and the *MCR*.

As a result, profits from "PGG Right" were significantly higher in the "exogenous institution" treatment (39.95) and in the "endogenous institution" treatment (36.95) compared to the "no institution" treatment (28.83) (Wilcoxon ranksum tests, p=0.000 and p=0.000). The difference between the "exogenous institution" treatment and the "endogenous institution" treatment is also significant (Wilcoxon ranksum test, p=0.000). Thus, the

[17] If not otherwise noted, reported results are from two-sided tests, based on part-II group averages.

institution was effective in raising cooperativeness and the increase in cooperation led to a large and significantly positive welfare effect for group members.

Spillover effects

In order to study the effects of an institution beyond its scope, we next analyze the contributions to "PGG Left" in the three treatments. As is visible in Figure 4.1 the contributions to "PGG Left" and "PGG Right" in the "no institution" treatment closely track each other and follow the typical declining pattern found in standard public goods games. Average contributions to "PGG Left" start at 11.45 in the first period and decline steadily to 4.33 in the last period and, averaged over all periods, there are no significant differences between contribution to "PGG Left" and "PGG Right" in the "no institution" treatment (Wilcoxon signed-rank tests, p=0.637). This shows that the subjects did not behave systematically different in the two games, when no institution was present.

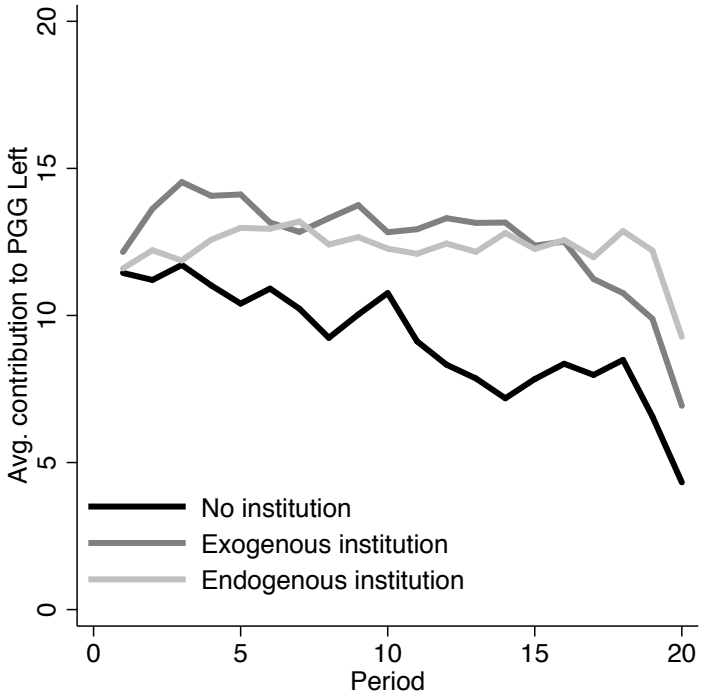


Figure 4.2: Contributions to "PGG Left".

In the "exogenous institution" treatment, average contributions to "PGG Left" start at 12.16 in the first period, reach a maximum of 14.53 in the third period, and thereafter decline steadily to 6.93 in the last period. In the "endogenous institution" treatment, on the other hand, average contributions to "PGG Left" start at 11.60 in the first period and remain in a corridor between 11.60 and 13.20 until the 20th period, in which they decline to 9.27. As can be seen from Figure 4.2, which displays average contributions

to "PGG Left", the contributions to "PGG Left" in the "exogenous institution" treatment and the "endogenous institution" treatment generally lie above the contributions to "PGG Left" in the "no institution" treatment. Over all periods, the average contributions to "PGG Left" are significantly higher in the "exogenous institution" treatment and the "endogenous institution" treatment compared to the "no institution" treatment (12.53 and 12.27 vs 9.15, respectively; Wilcoxon rank-sum tests, p=0.026 and p=0.047) and there is no significant difference between the "exogenous institution" treatment and "endogenous institution" treatment (Wilcoxon rank-sum tests, p=0.856). Thus, our first result is:

Result 1. *The presence of an institution that enforces cooperation in one domain leads to increased cooperation in other domains, beyond the reach of the institution.*

As a result of higher cooperation in "PGG Left", over all periods, the average profits from "PGG Left" are significantly higher in the "exogenous institution" treatment and the "endogenous institution" treatment compared to the "no institution" treatment (32.53 and 32.27 vs 29.15; Wilcoxon rank-sum tests, p=0.026 and p=0.047). Thus, institutions that regulate cooperation have significantly positive indirect welfare effects beyond their immediate effect in the regulated environment. The difference between the "exogenous institution" treatment and the "endogenous institution" treatment is not significant.

Table 4.2: Contributions to "PGG Left"

	Period				
	All	1-5	6-10	11-15	16-20
No Inst.	9.15	11.15	10.23	8.07	7.14
(N=23)	(1.07)	(.91)	(1.21)	(1.34)	(1.12)
Exo. Inst.	12.53	13.70	13.18	12.98	10.27
(N=22)	(.95)	(.83)	(1.17)	(1.28)	(1.12)
Endo. Inst.	12.27	12.25	12.70	12.35	11.78
(N=23)	(1.02)	(.84)	(1.13)	(1.32)	(1.31)
p-value (No vs. Exo.)	0.026**	0.047**	0.073*	0.013**	0.052*
p-value (No vs. Endo.)	0.047**	0.386	0.153	0.020**	0.013**
p-value (Exo. vs. Endo.)	0.856	0.166	0.716	1.000	0.427

Notes: Numbers in brackets are standard errors. p-values are from Wilcoxon rank-sum tests.

While aggregate levels of contributions do not differ between the "exogenous institution" treatment and the "endogenous institution" treatment, a closer look at Figure 4.2 suggests that the institution formation process can have an impact on the stability of the spillover. On average, contributions to "PGG Left" in the "exogenous institution" treatment still

decline over periods, while contributions to "PGG Left" in the "endogenous institution" treatment are almost perfectly flat until the last period. Indeed, while average contributions decline significantly over periods in both the "no institution" treatment and the "exogenous institution" treatment (Spearman's rho, ρ=-0.289, p=0.000 and ρ=-0.174, p=0.000, respectively), this is not the case in the "endogenous institution" treatment (Spearman rank order correlation, ρ=-0.022, p=0.639). Table 4.2 confirms that the advantage of the "endogenous institution" treatment emerges over time. In the first ten periods, there is no significant difference between the "no institution" treatment and the "endogenous institution" treatment. However, in the last 10 periods, a significant difference emerges and, due to the decay of contributions in the "exogenous institution" treatment, the contributions in the last 5 periods are highest in the "endogenous institution" treatment.

Result 2. *The institution formation process matters for the stability of the spillover in cooperative behavior. Exogenously imposed institutions do not lead to a stable contribution level beyond their scope, while endogenously adopted institutions do.*

These first two results are confirmed by regression analyses (see column (1) and (2) of Table 4.3). They show that, indeed, the exogenously imposed institution leads to a parallel upward shift of contribution levels, while the endogenously implemented institution leads to a change in the slope of contributions over time, thereby offsetting the typical decline in contributions.

In order to better understand the driving forces behind results 1 and 2, we next analyze, separately, the treatment effects on the decision to contribute and on the level of contribution (see Figure 4.3). Averaged over all periods, groups in the "exogenous institution" treatment show no significantly different proportion of subjects who contribute a positive amount than groups in the "no institution" treatment (78.18 percent vs 76.25 percent, Wilcoxon ranksum test, p=0.991). However, with 85.76 percent, groups in the "endogenous institution" treatment have a higher proportion of positive contributions and the difference to the "no institution" treatment and the "exogenous institution" treatment is marginally significant (Wilcoxon ranksum tests, p=0.098 and p=0.108, respectively). Thus, there exists some evidence that endogenously adopted institutions decrease free-riding behavior even in environments beyond their scope. If contributions are positive, they are significantly higher in the "exogenous institution" treatment (15.64) and the "endogenous institution" treatment (13.91) than in the "no institution" treatment (11.49) (Wilcoxon ranksum tests, p=0.001 and p=0.073, respectively). The difference between the "exogenous institution" treatment and the "endogenous institution" treatment is not significant (Wilcoxon ranksum test, p= 0.128).

Table 4.3: Treatment effect on contributions to "PGG Left"

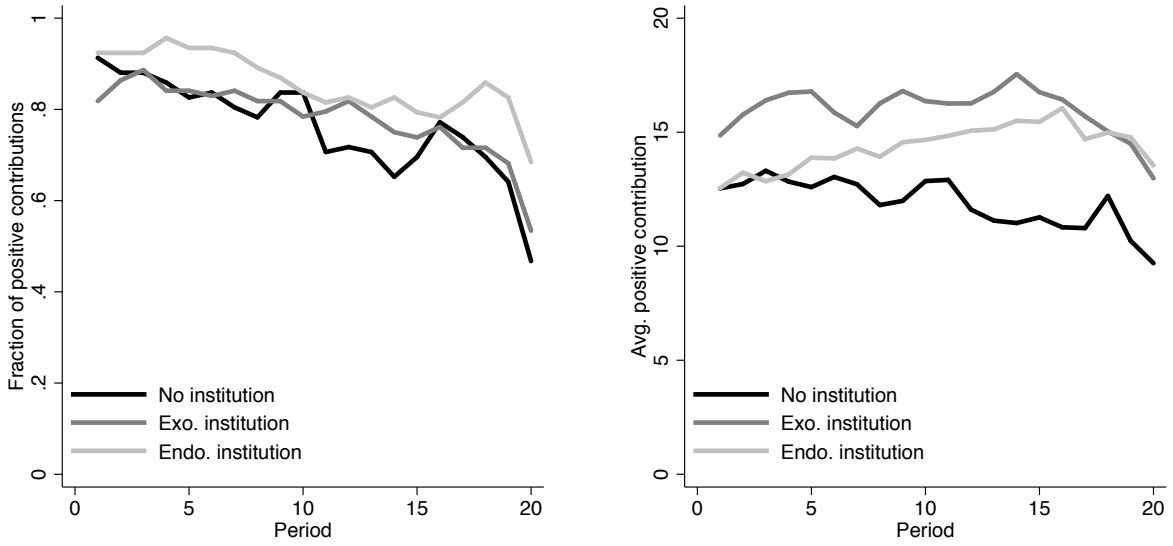
	(1) OLS Contributions	(2) Tobit panel Contributions
No (constant)	12.208***	13.911***
	(1.012)	(1.234)
Exo	2.540*	3.886**
	(1.405)	(1.771)
Endo	0.477	0.690
	(1.380)	(1.750)
Period	-0.291***	-0.490***
	(0.048)	(0.039)
Exo × Period	0.080	0.105*
	(0.088)	(0.056)
Endo × Period	0.252***	0.475***
	(0.091)	(0.055)
Observations	5440	5440
Adjusted R^2	0.062	

Notes: The independent variable is contributions to "PGG Left". The omitted category "No (constant)" is a binary variable that indicates participation in the "no institution" treatment. Robust standard errors (clustered on part-II groups in model (1)) in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The contribution dynamics in Panel 4.3b of Figure 4.3 reveal an intriguing pattern. The correlation between periods and contributions, when only positive contributions are considered, is significant and negative in the "no institution" treatment (ρ=-0.159, p=0.001), not significant in the "exogenous institution" treatment (ρ=-0.021, p=0.675), and significant and positive in the "endogenous institution" treatment (ρ=0.180, p=0.000). Result 3 summarizes these findings.[18]

Result 3. *Endogenously implemented institutions decrease free-riding behavior in domains beyond their scope compared to exogenously imposed institutions and to when no institution exists. Exogenously and endogenously implemented institutions have a positive and significant effect on the level of contributions beyond their scope compared to the "no institution" treatment.*

[18]These findings are mostly in line with the results from a 2-stage hurdle model that seperately estimates the treatment effect on the decision to contribute and the level of contribution (see Table A.1 in Appendix 4.A.2).



(a) Decision to contribute

(b) Contribution level

Figure 4.3: Decision to contribute and level of contribution

Next, we take a closer look at the relationship between the *MCR* and the contributions to "PGG Left". Taking group averages over all periods, contributions to "PGG Left" and the *MCR* in "PGG Right" are highly correlated (Spearman's rank order correlation, ρ=0.315, p=0.009).[19] However, such a positive effect of the *MCR* can be driven by a simple selection effect: those groups that consist of more cooperative subjects vote for a higher *MCR* and also contribute more to the public good. To test whether this happens, we focus on the two cases of *MCR*'s for which we have observations of randomly selected groups and groups that implemented the respective *MCR* endogenously. First, we only consider groups that implemented a *MCR* of zero and compare their contributions to groups in the "no institution" treatment (in which there is also an implicit *MCR* of zero). Subjects in groups with an endogenously implemented *MCR* of zero contribute, on average, 7.64 to "PGG Left", compared to 9.15 in the "no institution" treatment. This difference is not statistically significant (Wilcoxon ranksum test, p=0.572).[20] Similarly, the 19 groups in the "endogenous institution" treatment that implemented a *MCR* of 20 in some periods, contributed, on average, 12.34 to the "PGG Left" in those periods, compared to 12.53 in the "exogenous institution" treatment. Again, this difference is not statistically significant (Wilcoxon ranksum test, p=0.875). Hence, there is no difference in contributions between

[19]This result is confirmed in regression analysis (see column (1) of Table A.2 in Appendix 4.A.2). Column (2) shows that being in the "exogenous institution" treatment or the "endogenous institution" treatment compared to the "no institution" treatment has, if anything, a negative effect on contributions to "PGG Left" in addition to the effect through the *MCR*. However, the regression results have to be interpreted with caution as, in the "endogenous institution" treatment, behavior in past periods might both influence the implemented *MCR* and contributions to "PGG Left".

[20]To obtain the data for the "endogenous institution" treatment, we average a group's contribution to "PGG Left" over all periods in which that group implemented a *MCR* of zero. In total, only six groups implemented a *MCR* of zero at least once.

groups that endogenously implemented a *MCR* of zero or 20 and randomly selected groups. Thus, it seems that at least those groups which implemented the maximal and the minimal possible *MCR* are not different in their contribution behavior than randomly selected groups.[21]

Result 4. *Exogenously and endogenously implemented institutions increase cooperativeness beyond their scope through their effect on the* MCR*. There is no evidence that selection effects drive the results in the "endogenous institution" treatment.*

Voting behavior

This sections aims to shed some light on subjects' voting decisions. Figure 4.4 shows the fraction of votes for a *MCR* of 20, a *MCR* between 10 and 19, and a *MCR* of below 10 over the 20 periods. It is visible that subjects don't immediately vote for high *MCR*'s, but that votes for the highest possible *MCR* of 20 continuously increase and reach levels of over 80 percent after period 10. At the beginning, the largest fraction of subjects voted for a *MCR* between 10 and 19. As the fraction of votes for 20 increases, this fraction decreases continuously over time. Of the three groups, the fraction of subjects voting for a *MCR* below 10 is always the smallest and decreasing over time, but, until the end, there are, in every period, 3 to 4 subjects (out of a total of 92) who vote for such a low threshold. Due to the voting mechanism, which implements the minimal vote as the group's *MCR*, even when most subjects vote for 20, there can be a large fraction of groups with a *MCR* below 20. This is most apparent in the first period, where 17.39 percent of votes for a *MCR* of below 10 translate into 56.52 percent of groups with a *MCR* of below 10 and 28.26 percent of votes for a *MCR* of 20 translate into zero groups with a *MCR* of 20. In period 20, 89.13 percent of subjects vote for a *MCR* of 20 and 65.22 percent of groups implement it.

In the following, we study the determinants of the voting decision in period 1, before subjects had any experience with the mechanism and, thus, before their voting decision is endogenous to their experience in previous periods. In the first period of Part II, subjects vote, on average, for a *MCR* of 13.83 with the largest fractions voting for a *MCR* of 20 (28.26 percent), 15 (17.39 percent), and 10 (11.96 percent). There exists a significantly positive correlation between a subject's belief about his group members' contributions

[21] It could be that the *MCR* influences contributions in "PGG Left", because subjects simply imitate their contributions to "PGG Right" in "PGG Left". Fortunately, we have a good proxy to detect such behavior. Namely, subjects entered their contributions to "PGG Left" and "PGG Right" separately and we recorded the time at which each subject submitted their contribution decisions. Therefore, we know which decision they submitted first and how much time passed in between the two decisions. If subjects simply copied their contributions in "PGG Right" to "PGG Left", then this should hold especially for those who entered their decision for "PGG Right" first and who let little time pass in between their two decisions. However, we don't find this to be the case. For a detailed analysis, see Appendix 4.A.2.

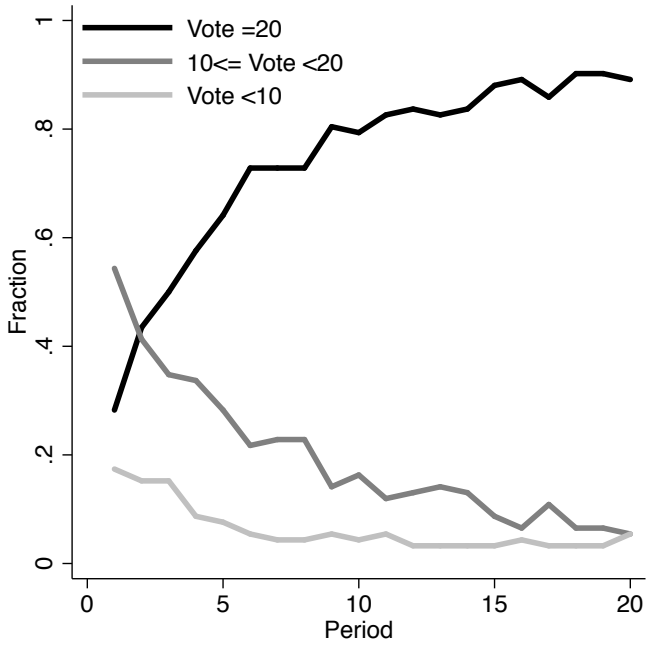


Figure 4.4: Voting behavior

in Part I and his vote in period 1. There also exists a significantly positive correlation between a subject's unconditional contribution decision in Part I and his vote in the first period of Part II. All other independent characteristics show no significant correlation (see Table A.4 in Appendix 4.A.2).[22] The decision to vote for any positive level of *MCR* is quite different from the decision to vote for the maximal possible *MCR* of 20 (only the latter can be interpreted as a signal that the subject wanted to maximize social efficiency). However, probit estimations for the effects of the individual characteristics on the likelihood to vote for a *MCR* of 20 no effect of any of the individual characteristics (see Table A.7 in Appendix 4.A.2).[23]

[22]OLS regressions of these individual characteristics on the vote in period 1 deliver a slightly different picture. Only a subject's unconditional contribution in Part I significantly predicts his vote in period 1 of Part II, but not his beliefs. Furthermore, there is a significantly positive effect of the stated level of altruism on the vote in period 1 (see A.5 in Appendix 4.A.2).

[23]The large amount of low votes in period 1 is puzzling as the guaranteed payoff from "PGG Right", when a *MCR* of 20 is implemented, is 40 ECU, whereas the actual average payoff from "PGG Right" in period 1 was 32.12 ECU and only 4 out of 92 subjects earned more than 40 ECU in period 1. It could be that subjects overestimate their group members' contributions and think they can earn more than 40 ECU by exploiting them. However, subjects' expected payoffs from "PGG Right", given their contributions and their (non-incentivized) beliefs about the contributions of their group members, is even significantly lower (31.14 ECU) than their actual payoff (Wilcoxon signed-rank test, p=0.072), and this holds also for the pivotal voters in period 1, who had no reason to update their beliefs between their vote and their contribution decision. Hence, these pivotal voters knowingly voted for a *MCR* that gave them a suboptimal payoff. Dal Bó, Dal Bó, and Eyster (2015) suggest that subjects underestimate the equilibrium effects of institutions which can lead to the demand for suboptimal policies. Whether this is also the case in our data and what drives subjects to pick up higher *MCR*'s over time lies beyond the scope of the present analysis.

4.4.2 Part I, III, IV - Effects on beliefs and preferences

Having established that there exists a significant spillover effect between domains that are regulated by an institutions and those that are not, we next analyze the persistent treatment effects on beliefs about others' cooperativeness, cooperative behavior and preferences for cooperation. The four panels of Figure 4.5 show, across parts and treatments, (a) the average midpoints of the stated belief intervals, (b) the average width of the stated belief intervals, (c) the average unconditional contributions, and (d) the average conditional contributions. The associated mean values, standard errors and econometric test results are summarized in Appendix 4.A.3, Tables A.11 to A.14. In Part I and for all four variables, there are no statistically significant differences across treatments.

Effects on beliefs about others' cooperativeness

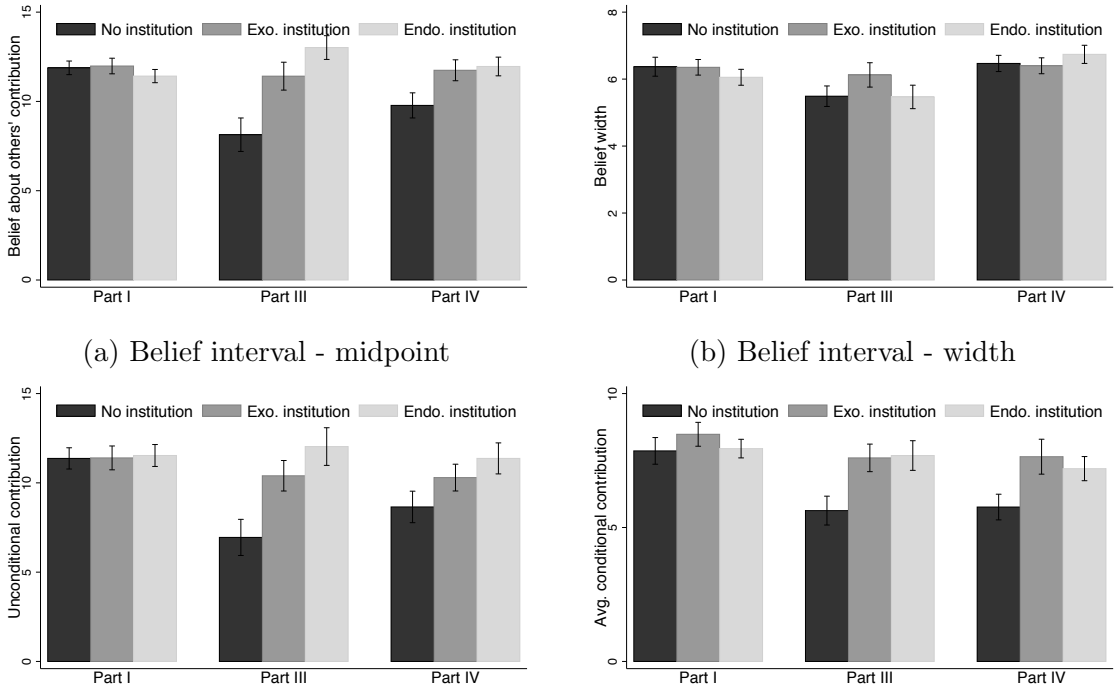
The data collected in the *belief elicitation stage* reveals that, in the "no institution" treatment, the midpoint of the provided belief interval is, on average, significantly lower in Part III (8.14) and IV (9.78) than in Part I (11.88) (Wilcoxon signed-rank test, p=0.000 and p=0.002). Thus, beliefs about others' cooperativeness decrease in the "no institution" treatment, but it decreases significantly more so for the group that one was associated with in Part II, than for a new group of strangers (Wilcoxon signed-rank test, p=0.005). In the "exogenous institution" treatment, the pattern is different. While the beliefs are a bit lower in Part III (11.41) and IV (11.74) than in Part I (11.98), the differences are not statistically significant. Thus, the exogenously imposed institution stopped the deterioration of beliefs. In the "endogenous institution" treatment, we find that beliefs about cooperativeness with regard to the part-II group members (Part III) is significantly higher than in Part I or Part IV (13.02 vs 11.42 and 11.97; Wilcoxon signed-rank test, p=0.018 and p=0.005). Furthermore, there is no significant difference between Part I and Part IV (Wilcoxon signed-rank test, p=0.260). Hence, the endogenously implemented institution has a positive effect on beliefs about own group members and keeps beliefs about strangers' cooperativeness stable.

Across treatments, we find that beliefs are significantly higher in the "exogenous institution" treatment and the "endogenous institution" treatment than in the "no institution" treatment in Part III (Wilcoxon rank-sum test, p=0.012 and p=0.000) and Part IV (Wilcoxon rank-sum test, p=0.047 and p=0.014). There is no significant difference between the two treatments with institution.[24]

[24]If we look at the very first period of Part II, when no history of play has been established and beliefs cannot be influenced by anything else than the institution, (non-incentivized) beliefs about others' contributions are highest in the "exogenous institution" treatment (12.43), and lower in the "no institution" treatment (11.34) and the "endogenous institution" treatment (11.5). However, only the difference

Result 5. *The presence of an institutions that enforces cooperation positively affects beliefs about the cooperativeness of group members even after the institution stopped existing. This extends to beliefs about cooperativeness of strangers, with whom there was no prior interaction.*

The average width of the belief interval in each part is not significantly different across treatments. In all treatments, however, the average width is smaller in Part III than in Part I. This difference is significant in the "no institution" treatment and the "endogenous institution" treatment (Wilcoxon signed-rank test, p=0.055 and p=0.032). On the other hand, the average width in Part IV is higher than in Part I in all treatments and significantly higher in the "endogenous institution" treatment (Wilcoxon signed-rank test, p=0.011). Thus, the repeated play over 20 periods increases subjects' confidence in their forecast of their group members' behavior, but this does not extend to strangers.



(a) Belief interval - midpoint

(b) Belief interval - width

(c) Unconditional contribution

(d) Avg. conditional contribution

Figure 4.5: Treatment effects on beliefs and cooperation.

Effects on pro-social behavior

Turning to the unconditional contributions in the three parts, a pattern similar to the one for beliefs emerges. In the "no institution" treatment, unconditional contributions

between the "exogenous institution" treatment and the "no institution" treatment is significant at the 10 percent level (Wilcoxon ranksum test, p=0.097).

drop significantly from, on average, 11.37 in Part I to 6.95 in Part III (Wilcoxon signed-rank test, p= 0.000) and 8.65 in Part IV (Wilcoxon signed-rank test, p=0.002). In the "exogenous institution" treatment and the "endogenous institution" treatment, on the other hand, unconditional contributions remain stable, on average, with no significant differences between Part I, Part III and Part IV. Across treatments, this leads to significantly higher unconditional contribution levels in the "exogenous institution" treatment and the "endogenous institution" treatment compared to the "no institution" treatment in Part III (Wilcoxon rank-sum tests, p=0.009 and p=0.002) and significantly higher contributions in the "endogenous institution" treatment compared to the "no institution" treatment in Part IV (Wilcoxon rank-sum tests, p=0.017).[25]

Result 6. *The presence of an institutions that enforces cooperation increases cooperative behavior even after the institution stopped existing. This extends to cooperative behavior towards strangers, with whom there was no prior interaction.*

Effects on preferences for cooperation

Importantly, institutions do not only have an effect on cooperative behavior and beliefs about others' cooperativeness, but also on preferences for cooperation. We find that the average conditional contribution, i.e. the average amount a subject decided to contribute conditional on all possible contributions of others (see Figure 4.5d), in Part III and Part IV is significantly higher in the "exogenous institution" treatment (7.60 in Part III, 7.64 in Part IV) and the "endogenous institution" treatment (7.69 in Part III, 7.20 in Part IV) than in the "no institution" treatment (5.63 in Part III, 5.76 in Part IV) (Wilcoxon rank-sum tests, p=0.005 and p=0.008 in Part III, p=0.011 and p=0.024 in Part IV). As with beliefs and unconditional contributions, this difference is mainly driven by a deterioration of preferences for cooperation in the "no institution" treatment from Part I to Part III and IV which does not happen in the "exogenous institution" treatment and the "endogenous institution" treatment. Table 4.4 confirms the last four results in regression analyses.

[25]The finding that contributions in the "endogenous institution" treatment are significantly higher than in the "no institution" treatment also provides some evidence that contributions in the "PGG Left" of Part II were not higher in the "endogenous institution" treatment because group members were afraid of loosing the institution in "PGG Right" in the next period. In Part III and IV there is no more institution to implement and contributions are still significantly higher in the "endogenous institution" treatment. Furthermore, it also provides evidence against a risk channel. In Part II, if agent's are uncertain about their group members' contributions, contributing to the public good becomes a risky choice for conditional cooperators. If they have decreasing absolute risk aversion they will contribute/invest more when they receive a sure income in "PGG Right". This channel cannot explain the findings in Part III and IV in which there is no such difference in guaranteed income across treatments (however, subjects could still be influenced by their, in expectation, higher income from Part II).

Table 4.4: Treatment effects on belief midpoint, belief width, unconditional contribution and average conditional contribution.

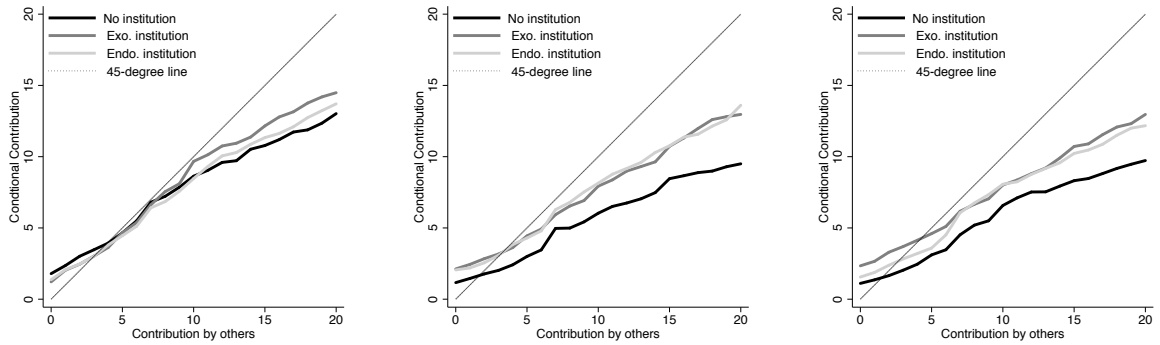
	(1) OLS Belief midpoint	(2) OLS Belief width	(3) OLS Unconditional contribution	(4) OLS Avg. conditional contribution
Part I (constant)	11.880***	6.370***	11.370***	7.860***
	(0.378)	(0.281)	(0.590)	(0.493)
Exo	0.103	-0.017	0.028	0.620
	(0.574)	(0.363)	(0.886)	(0.661)
Endo	-0.462	-0.315	0.163	0.086
	(0.527)	(0.366)	(0.847)	(0.601)
Part III	-3.745***	-0.880**	-4.424***	-2.229***
	(0.701)	(0.406)	(0.873)	(0.426)
Part III × Exo	3.176***	0.653	3.424***	1.348*
	(1.045)	(0.569)	(1.268)	(0.681)
Part III × Endo	5.342***	0.293	4.924***	1.969***
	(0.927)	(0.513)	(1.235)	(0.689)
Part IV	-2.103***	0.098	-2.717***	-2.096***
	(0.508)	(0.254)	(0.722)	(0.343)
Part IV × Exo	1.865**	-0.052	1.615*	1.260*
	(0.713)	(0.373)	(0.954)	(0.659)
Part IV × Endo	2.641***	0.587*	2.554**	1.348**
	(0.699)	(0.342)	(1.012)	(0.561)
Observations	816	816	816	816
Adjusted R^2	0.083	0.016	0.046	0.026

Notes: The omitted category "Part I (constant)" is a binary variable indicating the decision was made in Part I. Robust standard errors (clustered on part-II groups) in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Result 7. *The presence of an institutions that enforces cooperation increases preferences for cooperation even after the institution ceases to exist. This extends to preferences for cooperation towards strangers, with whom there was no prior interaction.*

In the following, we take a closer look at the treatment effects on conditional contributions and thus prosocial preferences. The three panels of Figure 4.6 show the average conditional contribution levels for each possible average contribution by others in the three parts. As can be seen from the figure, in Part III and IV the conditional contributions in the "no institution" treatment lie considerably below the level of conditional contributions in the

"exogenous institution" treatment and the "endogenous institution" treatment.



(a) Part I (b) Part III (c) Part IV

Figure 4.6: Conditional contribution

Two different preference changes could lead to an overall positive effect of institution on preferences for cooperation compared to when no institution was present. First, institutions could change preferences for pro-social behavior which would lead to a constant increase or decrease in contributions for all possible average contribution levels of the other group members, thus affecting the intercept of the conditional contribution schedule. Second, institutions could also affect preferences for reciprocation which would show up as a change in the slope of conditional contribution decisions. For example, this would be the case if a subjects who experienced the institution increases his contributions when his group members contribute a lot, but not when his group members contribute little, compared to a subject who did not experience the institution.

In order to disentangle these two potential effects, we regress the treatment, the parts, the contributions by others, and the interactions between the variables on the conditional contribution decisions (see Table A.15 in Appendix 4.A.3). The regression confirms that contribution schedules are significantly increasing in the other group members' contributions and that, in Part I, there are no significant differences between the three treatments with the exception of the slope of the contribution schedule in the "exogenous institution" treatment, which is marginally significantly steeper than in the "no institution" treatment. In both, the "exogenous institution" treatment and the "endogenous institution" treatment, the intercept of the contribution schedules is significantly higher than in the "no institution" treatment in parts III and IV. The slope, however, is not significantly different in the "exogenous institution" treatment and the "endogenous institution" treatment compared to the "no institution" treatment in parts III and IV. Thus, these findings indicate that the institution changes preferences for pro-social behavior and not preferences for reciprocation and that this change in preferences for pro-social behavior extends to behavior directed at strangers, with whom a subject never interacted before.

Result 8. *The presence of an institutions that enforces cooperation increases preferences*

for prosocial behavior, but leaves reciprocal preferences unchanged.

4.5 Discussion

Next, we discuss our results with respect to the different hypotheses suggested in Section 2. To begin with, the data clearly rejects the null hypothesis of no institutional influence on cooperative behavior in the "PGG Left" of Part II. In both, the "exogenous institution" treatment and the "endogenous institution" treatment, contribution levels in "PGG Left" are significantly higher than in the "no institution" treatment. Furthermore, as documented in Part III and IV, cooperative behavior, preferences for cooperation and beliefs about others' cooperativeness are significantly higher in the "endogenous institution" treatment and the "exogenous institution" treatment. This confirms our hypothesis 2-4.

In the following, we try to shed light on which of the proposed channels drive the results for cooperative preferences and beliefs. First, since we don't find a crowding-out of intrinsic preferences for cooperation, the proposed *pure altruism* channel, the *control aversion* channel in the "exogenous institution" treatment and the *distrust aversion* channel in the "endogenous institution" treatment can be dismissed as not being dominant in affecting preferences for cooperation and cooperative behavior in our setting.

Second, one potential preference channel was that subjects follow simple rules that are provided to them and that the institution in the "exogenous institution" treatment and the "endogenous institution" treatment provides such a rule. Our measure of rule-following propensity (RFT), which was elicited in Part V, allows us to detect such an effect. If the institution is perceived as a rule that one should follow, then subjects who are more willing to follow a randomly chosen rule in the RFT, should, ceteris paribus, also be more willing to follow the social rule highlighted by the institution in the "exogenous institution" treatment and "endogenous institution" treatment (cf. Kimbrough and Vostroknutov (forthcoming)). However, we find no significant effects of the interactions between subjects' rule-following propensity and the "exogenous institution" treatment or the "endogenous institution" treatment on contributions to "PGG Left", (see column (1) of Table A.9 in Appendix 4.A.2). Thus, there is no evidence that subjects interpret the institution as a rule that they have to follow.

Third, another potential preference channels is provided by the Social Heuristics Hypotheses (SHH) which proposes that cooperative norms establish simple heuristics that are followed intuitively. We can test whether channel is active using the results from the Cognitive Reflection Test (CRT), which was elicited in Part V. The CRT measures

how intuitively subjects answer questions, with a lower score meaning higher reliance on intuition. Therefore, if the institution in the "exogenous institution" treatment and the "endogenous institution" treatment provides a heuristic for cooperation, then, according to the SHH, subjects who score lower on the CRT should follow the norm intuitively and contribute more in those treatments (cf. the argument in Peysakhovich and Rand (forthcoming), chapter 3.2.3). Indeed, we find a negative effect of the interaction between the CRT score and the "exogenous institution" treatment and the "endogenous institution" treatment on contributions to "PGG Left". This effect is significant for the "exogenous institution" treatment (see column (2) of Table A.9 in Appendix 4.A.2). Therefore, we find evidence that the exogenously implemented institution provides a cooperative norm which is followed intuitively.

Fourth, we argued that institutions could affect beliefs about group member's cooperativeness and that this influences contributions of conditional cooperators. Part III and IV demonstrate that institutions indeed affect beliefs positively compared to the case without an institution. More specifically, in the "endogenous institution" treatment, beliefs could be affected if the others' votes and thus the implemented institution provides a valuable signal about their cooperative type. However, as suggested, the effect could go both ways. Higher implemented *MCR*'s could signal that group members are selfish and cooperation needs to be enforced, or that group members understood the game and agreed that high contributions are what the group should aim for. While beliefs about others' cooperativeness are higher in the "endogenous institution" treatment than in the "no institution" treatment in both, Part III and IV, we also find that, within the "endogenous institution" treatment, beliefs in Part III are significantly higher than in Part I, before the institution was implemented, and also significantly higher than in Part IV, when subjects played with new group members. The same patter is not found in the "exogenous institution" treatment where beliefs are not significantly different across parts. Thus, this can be interpreted as evidence that the process of endogenous implementation provides subjects with a signal that is specific to their group members' cooperativeness and that is not informative about cooperativeness of strangers. Indeed, when only considering period 1 of Part II, subjects' non-incentivized beliefs about the average contribution of their group members show a positive and significant correlation with their group members' average vote for the *MCR* (Spearman's rho, ρ=0.306, p=0.003). This indicates that the vote is interpreted as a positive signal about others' cooperativeness.[26]

To summarize, we find some evidence that institutions induce a behavioral spillover to non-monitored domains through their effect on beliefs and by providing an intuitive heuristic that people follow. We don't find evidence that subjects interpret the institution like as

[26] A regression analysis confirms that this result is robust to the inclusion of the implemented *MCR* and a subject's own vote as control variables (see Table A.10 in Appendix 4.A.2).

a rule that they should follow. However, effects on beliefs and preferences also reinforce each other, which makes it difficult to disentangle the relative strength of the different channels.[27]

4.6 Conclusion

In this paper, we experimentally demonstrate that simple institutions which regulate cooperative behavior affect behavior beyond their scope. We also document an interesting difference between exogenously imposed and endogenously adopted institutions. The exogenously imposed institution leads to an immediate positive impact on voluntary cooperation, but the decline in cooperation over time is similar to the case without an institution. The endogenous institution, on the other hand, produces a weaker immediate effect, which was partly due to adoption of weaker institution in the beginning, but the effect perseveres more strongly than that of the exogenous institution. We also find that both types of institutions have effects that persist beyond their presence. Treated subjects have more positive beliefs about others' contributions and they contribute more, both conditionally and unconditionally. These effects also extend to new counterparts, with whom no previous interaction occurred.

Beyond any immediate effects of institutions on pro-social behavior, understanding "spillover effects" is important. Our results demonstrate that regulating cooperation does not necessarily crowd out intrinsic motivation to cooperate. In fact, the opposite holds true. Preferences for pro-social behavior are persistently higher even after an institution ceases to exist compared to when such an institution never existed. This has important consequences for policy making as regulators can hope for sizable spillover effects that have to be taken into account when deciding about the introduction of new regulations and policies.

From a broader perspective, our study can also speak to the literature on the interrelation between institutions and culture (Guiso, Sapienza, and Zingales, 2006, 2008; Tabellini, 2008, 2010; Alesina and Guiliano, forthcoming). The set of beliefs and preferences that the members of a society hold are commonly acknowledged as important determinants of a society's culture. With this regard, we provide causal evidence that institutions can shape culture persistently, and that the institutionally induced change in culture can lead to sizable welfare effects.

[27] There is a highly significant positive correlation in Part I between beliefs about others' cooperativeness and preferences for cooperation as measured by the average conditional contribution (Spearman's rho, ρ=0.353, p=0.000).

Of course, more work is necessary to understand when and how the behavioral effects of institutions extend to unregulated behaviors and settings. In principle, many different institutions and settings are conceivable. For example, while our exogenously implemented institution costlessly enforced the socially most efficient cooperation level, many institutions in the real world come with some form of inefficiency or related costs. Future research should therefore examine how robust our findings are to the variation of the institutions and whether different institutions can lead to different, interesting effects. Furthermore, future work should shed light on how institutions that regulate behavior in one environment spill over to influence different kinds of behavior in other, unregulated environments. For example, an interesting question could be how an institution that fosters competitive behavior in one setting affects behavior in another setting that relies on cooperation, or vice versa. The present study should therefore be seen as the starting point of a new area of research that can extend in many promising directions.

Bibliography

Alesina, A., and P. Guiliano (forthcoming): "Culture and Institutions," *Journal of Economic Literature*.

Andreoni, J. (1988): "Privately provided public goods in a large economy: The limits of altruism," *Journal of Public Economics*, 35(1), 57–73.

——— (1990): "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving," *The Economic Journal*, 100(401), 464–477.

Becker, A., T. Deckers, T. Dohmen, A. Falk, and F. Kosse (2012): "The Relationship Between Economic Preferences and Psychological Personality Measures," *Annual Review of Economics*, 4(1), 453–478.

Bednar, J., Y. Chen, T. X. Liu, and S. Page (2012): "Behavioral spillovers and cognitive load in multiple games: An experimental study," *Games and Economic Behavior*, 74(1), 12–31.

Belot, M., and M. Schröder (2015): "The Spillover Effects of Monitoring: A Field Experiment," *Management Science*, (April), 150313071440008.

Benabou, R., and J. Tirole (2011, WP): "Laws and Norms," *NBER Working Paper Series*.

Bernasconi, M., L. Corazzini, S. Kube, and M. A. Maréchal (2009): "Two are better than one! Individuals' contributions to "unpacked" public goods," *Economics Letters*, 104(1), 31–33.

Bernheim, B. D. (1986): "On the voluntary and involuntary provision of public goods," *American Economic Review*, 151(3712), 789–793.

Blackwell, C., and M. McKee (2003): "Only for my own neighborhood? Preferences and voluntary provision of local and global public goods," *Journal of Economic Behavior & Organization*, 52(1), 115–131.

Bock, O., I. Baetge, and A. Nicklisch (2014): "hroot: Hamburg Registration and Organization Online Tool," *European Economic Review*, 71, 117 – 120.

Bowles, S., and S. Polanía-Reyes (2012): "Economic Incentives and Social Preferences: Substitutes or Complements?," *Journal of Economic Literature*, 50(2), 368–425.

Brandts, J., and D. J. Cooper (2006): "A Change Would Do You Good .... An Experimental Study on How to Overcome Coordination Failure in Organizations," *American Economic Review*, 96(3), 669–693.

Cason, T. N., and L. Gangadharan (2013): "Cooperation Spillovers and Price Competition in Experimental Markets," *Economic Inquiry*, 51(3), 1715–1730.

Cason, T. N., A. C. Savikhin, and R. M. Sheremeta (2012): "Behavioral spillovers in coordination games," *European Economic Review*, 56(2), 233–245.

Cassar, A., G. D'Adda, and P. Grosjean (2014): "Institutional Quality, Culture, and Norms of Cooperation: Evidence from a Behavioral Field Experiment," *Journal of Law and Economics*, 57(3), 821–863.

Cooter, R. (1998): "Expressive Law And Economics," *The Journal of Legal Studies*, 27(S2), 585–607.

Dal Bó, E., P. Dal Bó, and E. Eyster (2015): "The Demand for Bad Policy when Voters Underappreciate Equilibrium Effects," *Working Paper*, pp. 1–47.

Dal Bó, P., A. Foster, and L. Putterman (2010): "Institutions and Behavior: Experimental Evidence on the Effects of Democracy," *American Economic Review*, 100(5), 2205–2229.

Deci, E. L. (1975): *Intrinsic Motivation*. Plenum Press, New York.

Deci, E. L., R. Koestner, and R. M. Ryan (1999): "A Meta-Analyitc Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation," *Psychological Bulletin*, 125(6), 627–628.

Deci, E. L., and R. M. Ryan (1985): *Intrinsic Motivation and Self-Determination in Human Behavior*. Plenum Press, New York.

Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp, and G. G. Wagner (2011): "Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences," *Journal of the European Economic Association*, 9(3), 522–550.

Faillo, M., D. Grieco, and L. Zarri (2013): "Legitimate punishment, feedback, and the enforcement of cooperation," *Games and Economic Behavior*, 77(1), 271–283.

Falk, A., E. Fehr, and C. Zehnder (2006): "Fairness Perceptions and Reservation Wages - The Behavioral Effects of Minimum Wage Laws," *The Quarterly Journal of Economics*, 121(4), 1347–1381.

Falk, A., U. Fischbacher, and S. Gächter (2013): "Living in Two Neighborhoods-Social Interaction Effects in the Laboratory," *Economic Inquiry*, 51(1), 563–578.

Falk, A., and M. Kosfeld (2006): "The Hidden Costs of Control," *The American economic review*, 96(December), 1611–1630.

Fehr, E., and U. Fischbacher (2004): "Social norms and human cooperation.," *Trends in cognitive sciences*, 8(4), 185–90.

Fehr, E., U. Fischbacher, and S. Gächter (2002): "Strong reciprocity, human cooperation, and the enforcement of social norms," *Human Nature*, 13(1), 1–25.

Fehr, E., and K. M. Schmidt (1999): "A Theory of Fairness, Competition, and Cooperation," *The Quarterly Journal of Economics*, 114(3), 817–868.

Fellner, G., and G. K. Lünser (2014): "Cooperation in local and global groups," *Journal of Economic Behavior & Organization*, pp. 1–10.

Fischbacher, U. (2007): "z-Tree: Zurich toolbox for ready-made economic experiments," *Experimental Economics*, 10(2), 171–178.

Fischbacher, U., and S. Gächter (2010): "Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods Experiments," *American Economic Review*, 100(1), 541–556.

Fischbacher, U., S. Gächter, and E. Fehr (2001): "Are people conditionally cooperative ? Evidence from a public goods experiment," *Economics Letters*, 71(3), 397–404.

Frederick, S. (2005): "Cognitive Reflection and Decision Making," *The Journal of Economic Perspectives*, 19(4), 25–42.

Frey, B. S. (1993): "Motivation as a limit to pricing," *Journal of Economic Psychology*, 14(4), 635 – 664.

Frey, B. S. (1997): "A Constitution for Knaves Crowds Out Civic Virtues," *The Economic Journal*, 107(443), 1043–1053.

Frey, B. S., and R. Jegen (2001): "Motivation Crowding Theory," *Journal of Economic Surveys*, 15(5), 589–611.

Funk, P. (2007): "Is There An Expressive Function of Law? An Empirical Analysis of Voting Laws with Symbolic Fines," *American Law and Economics Review*, 9(1), 135–159.

Gächter, S., and E. Fehr (2000): "Cooperation and Punishment in Public Goods Experiments," *American Economic Review*, 90(4), 980–994.

Galbiati, R., K. H. Schlag, and J. J. van der Weele (2013): "Sanctions that signal: An experiment," *Journal of Economic Behavior & Organization*, 94, 34–51.

Galbiati, R., and P. Vertova (2008): "Obligations and cooperative behaviour in public good games," *Games and Economic Behavior*, 64(1), 146–170.

Gneezy, U., S. Meier, and P. Rey-Biel (2011): "When and Why Incentives (Don't) Work to Modify Behavior," *Journal of Economic Perspectives*, 25(4), 191–210.

Greiner, B. (2003): "An Online Recruitment System for Economic Experiments," in *Forschung und wissenschaftliches Rechnen 2003. GWD Bericht 62*, ed. by K. Kremer, and V. Macho, pp. 79–93. Ges. für Wiss. Datenverarbeitung, Göttingen.

Grimm, V., and F. Mengel (2012): "An experiment on learning in a multiple games environment," *Journal of Economic Theory*, 147(6), 2220–2259.

Guiso, L., P. Sapienza, and L. Zingales (2006): "Does Culture Affect Economic Outcomes?," *Journal of Economic Perspectives*, 20(2), 23–48.

——— (2008): "Alfred Marshall Lecture Social Capital As Good Culture," *Journal of the European Economic Association*, 6(2-3), 295–320.

Gürerk, O., B. Irlenbusch, and B. Rockenbach (2006): "The competitive advantage of sanctioning institutions.," *Science (New York, N.Y.)*, 312(5770), 108–11.

Herz, H., and D. Taubinsky (2013, WP): "Market Experience is a Reference Point in Judgments of Fairness," *Working Paper*.

Huck, S., P. Jehiel, and T. Rutter (2011): "Feedback spillover and analogy-based expectations: A multi-game experiment," *Games and Economic Behavior*, 71(2), 351–365.

Irlenbusch, B., and G. K. Ruchala (2008): "Relative rewards within team-based compensation," *Labour Economics*, 15(2), 141 – 167.

Jehiel, P. (2005): "Analogy-based expectation equilibrium," *Journal of Economic Theory*, 123(2), 81–104.

Kahan, D. M. (1998): "Social Meaning and the Economic Analysis of Crime," *The Journal of Legal Studies*, 27(S2), 609–622.

Kamei, K. (2014, WP): "Democracy and Resilient Pro-Social Behavioral Change: An Experimental Study," *Working Paper*.

Kimbrough, E. O., J. Miller, and A. Vostroknutov (2014): "Norm-Dependent Utility in Games," *Working Paper*, (June), 1–26.

Kimbrough, E. O., and A. Vostroknutov (forthcoming): "Norms Make Preferences Social," *Journal of the European Economic Association*.

Knez, M., and C. Camerer (2000): "Increasing Cooperation in Prisoner's Dilemmas by Establishing a Precedent of Efficiency in Coordination Games.," *Organizational behavior and human decision processes*, 82(2), 194–216.

Kosfeld, M., A. Okada, and A. Riedl (2009): "Institution Formation in Public Goods Games," *American Economic Review*, 99(4), 1335–1355.

Kroll, S., T. L. Cherry, and J. F. Shogren (2007): "Voting, Punishment, and Public Goods," *Economic Inquiry*, 45(3), 557–570.

Krupka, E. L., and R. a. Weber (2013): "Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary?," *Journal of the European Economic Association*, 11(3), 495–524.

Markussen, T., L. Putterman, and J.-R. Tyran (2014): "Self-Organization for Collective Action: An Experimental Study of Voting on Sanction Regimes," *The Review of Economic Studies*, 81(1), 301–324.

Masclet, D., C. Noussair, S. Tucker, and M.-C. Villeval (2003): "Monetary and Nonmonetary Punishment in the Voluntary Contributions Mechanism," *American Economic Review*, 93(1), 366–380.

McCarter, M. W., A. C. Samak, and R. M. Sheremeta (2013, WP): "Divided Loyalties or Conditional Cooperation? An Experimental Study of Contributions to Multiple Public Goods," *Working Paper*.

Ostrom, E., J. Walker, and R. Gardner (1992): "Covenants With and Without a Sword: Self-Governance is Possible," *The American Political Science Review*, 86(2), 404–417.

Peysakhovich, A., and D. Rand (forthcoming): "Habits of virtue: creating norms of cooperation and defection in the laboratory," *Management Science*, pp. 1–51.

Putterman, L., J.-R. Tyran, and K. Kamei (2011): "Public goods and voting on formal sanction schemes," *Journal of Public Economics*, 95(9-10), 1213–1222.

Rand, D. G., A. Peysakhovich, G. T. Kraft-Todd, G. E. Newman, O. Wurzbacher, M. a. Nowak, and J. D. Greene (2014): "Social heuristics shape intuitive cooperation.," *Nature communications*, 5, 3677.

Reuben, E., and A. Riedl (2013): "Enforcement of contribution norms in public good games with heterogeneous populations," *Games and Economic Behavior*, 77(1), 122–137.

Savikhin, A. C., and R. M. Sheremeta (2013): "Simultaneous Decision-Making in Competitive and Cooperative Environments," *Economic Inquiry*, 51(2), 1311–1323.

Schlag, K. H., and J. van der Weele (2012): "Incentives for Eliciting Confidence Intervals," *Working Paper*, December, 1–24.

Sliwka, D. (2007): "Trust as a Signal of a Social Norm and the Hidden Costs of Incentive Schemes," *American Economic Review*, 97(3), 999–1012.

Sunstein, C. (1996): "On the Expressive Function of Law," *University of Pennsylvania Law Review*, 144(5), 2021–2053.

Sutter, M., S. Haigner, and M. G. Kocher (2010): "Choosing the Carrot or the Stick? Endogenous Institutional Choice in Social Dilemma Situations," *Review of Economic Studies*, 77(4), 1540–1566.

Tabellini, G. (2008): "The Scope of Cooperation: Values and Incentives," *The Quarterly Journal of Economics*, 123(3), 905–950.

——— (2010): "Culture and Institutions: Economic Development in the Regions of Europe," *Journal of the European Economic Association*, 8(4), 677–716.

Tausch, F. (2014): "The benefits of external control," *Working Paper*.

Tyran, J.-R., and L. P. Feld (2006): "Achieving Compliance when Legal Sanctions are Non-deterrent," *Scandinavian Journal of Economics*, 108(1), 135–156.

van der Weele, J. (2009): "The Signaling Power of Sanctions in Social Dilemmas," *Journal of Law, Economics, and Organization*, 28(1), 103–126.

Vischer, T., T. Dohmen, A. Falk, D. Huffman, J. Schupp, U. Sunde, and G. G. Wagner (2013): "Validating an ultra-short survey measure of patience," *Economics Letters*, 120(2), 142 – 145.

Walker, J. M., R. Gardner, A. Herr, and E. Ostrom (2000): "Collective Choice in the Commons: Experimental Results on Proposed Allocation Rules and Votes," *The Economic Journal*, 110(460), 212–234.

4.A Appendix

4.A.1 Theoretical predictions

Standard preferences

If we assume that players are only motivated to maximize their own material payoff, the game-theoretic predictions for the one-stage game are as follows. Since the material payoff from the "PGG Left" is independent of the "PGG Right", and vice versa, a player's overall utility U_i is additively separable into a utility-part from "PGG Left", u_i^L, and one from "PGG Right", u_i^R.

Given the contributions of all other players, the utility u_i^L of player i from "PGG Left" is equal to

$$u_i^L(g_1^L, ..., g_4^L) = w^L - g_i^L + a\sum_{j=1}^{4} g_j^L, \tag{4.3}$$

where $0 < a < 1 < 4a$. The parameter a models the marginal per capita return (MPCR) from contributing to the public good, w^L is the per period endowment, and g_i^L is player i's contribution to the public good in "PGG Left". Assumption $a < 1$ implies that contributing nothing is the strictly dominant action for every player with standard preferences because every player's material payoff is maximized by contributing zero to the public good regardless of the other players' contributions. In consequence, the strategy profile $(0, 0, 0, 0)$ is the unique Nash equilibrium of the "PGG Left".

The utility u_i^R of player i from "PGG Right" is equal to

$$u_i^R(g_1^R, ..., g_4^R) = \begin{cases} w^R - g_i^R + a\sum_{j=1}^{4} g_j^R & \text{if } g_i^R \geq MCR \\ 0 & \text{if } g_i^R < MCR, \end{cases} \tag{4.4}$$

where MCR is the minimum contribution requirement that is implemented by the institution. The MCR is equal to zero in the "no institution" treatment, equal to w^R in the "exogenous institution" treatment, and equal to the outcome of the voting process, in the "endogenous institution" treatment.

Because the institution deters any material incentive to contribute less than the contribution threshold, the dominant action for every player with standard preferences is to contribute MCR in "PGG Right". In the "no institution" treatment, the strategy profile $\{(0, 0), (0, 0), (0, 0), (0, 0)\})$ is thus the unique Nash equilibrium of the entire game. The numbers in brackets stand for the contribution to "PGG Left" (left number) and "PGG

Right" (right number). In the "exogenous institution" treatment, the strategy profile $\{(0, w^R), (0, w^R), (0, w^R), (0, w^R)\}$ is the unique Nash equilibrium.

In the "endogenous institution" treatment, every period consists of two stages - a voting stage and a contribution stage. Therefore both the voting behavior in the first stage as well as the later contributions are part of a player's optimizing strategy. In a subgame perfect equilibrium, players decide on their actions in every stage, rationally anticipating the outcome of future stages by applying backward induction. Consider first the contribution stage. If players attempt to maximize their material payoff, they contribute nothing in the "PGG Left" and *MCR* in "PGG Right". Given this behavior in the contribution stage one can derive the optimal voting behavior in the voting stage. The optimal vote v_i^* depends on the other players' votes $\mathbf{v}_{-i}$:

$$v_i^* \in [\min\{\mathbf{v}_{-i}\}, w^R]$$

Assumption $4a > 1$ implies that all players are better off if everyone contributes his full endowment to the public good. Therefore, voting $v_i = w^R$ is a weakly dominant strategy: If all other players voted w^R, player i is strictly better off by voting $v_i = w^R$ as well. If at least one player voted less than w^R, player i is indifferent between voting for a threshold in $[\min\{\mathbf{v}_{-i}\}, w^R]$.

PROPOSITION 1: *If players have standard preference, there exists a unique strict subgame perfect equilibrium in which all players vote to set the* MCR *equal to the full endowment ($v_i = w^R \; \forall \; i$), contribute their full endowment in "PGG Right" ($g_i^R = w^R \; \forall \; i$), and contribute nothing in "PGG Left" ($g_i = 0 \; \forall \; i$).*[28]

[28]Note that there exist other subgame perfect equilibria in which at least two players vote for the same *MCR* $v_i = v_j < w^R$, $v_i = \min\{(v_i)_{i \in I}\}$, all contribute $v_i = MCR$ in "PGG Right" and 0 in "PGG Left". This is the case because deviating in their vote from v_i would not change the implemented threshold (only the smallest vote counts).

4.A.2 Part II - Results

Treatment effects on the decision to contribute and the level of contribution

Table A.1 shows the two stages of a hurdle model that estimates, separately, the treatment effect on the decision to contribute and the level of contribution. In the first stage, a Probit regression estimates the effect of treatments, period, and treatment-period interactions on the decision to contribute. Only the period has a significant (and negative) effect on the decision to contribute. Both, the "exogenous institution" treatment and the "endogenous institution" treatment don't influence the decision to contribute significantly, compared to the "no institution" treatment. They also do not alter the dynamics. Over time, more and more group members decide not to contribute at all and this effect is not statistically different in the "exogenous institution" treatment and the "endogenous institution" treatment compared to the "no institution" treatment. The second stage of the hurdle model estimates a linear regression model truncated at zero. Hence, it measures the effect on the level of contribution, if a subject decided to contribute a positive amount. Here, we do find that the "exogenous institution" treatment leads to significantly higher contribution levels compared to the "no institution" treatment, while there is no such effect for the "endogenous institution" treatment. On the other hand, we again find that later periods lead to significantly lower contribution rates and that this effect is not attenuated when subjects are in the "exogenous institution" treatment. In the "endogenous institution" treatment, however, the trend is significantly different from the "no institution" treatment. In fact, in the "endogenous institution" treatment, the decline over time is more than offset, such that, for those who decide to contribute, contribution levels rise over time.

Table A.1: Contributions to "PGG Left"

	(1) Hurdle Stage 1	(2) Hurdle Stage 2
No (constant)	1.294***	13.213***
	(0.193)	(0.921)
Exo	-0.087	3.142***
	(0.285)	(1.069)
Endo	0.298	-0.341
	(0.290)	(1.242)
Period	-0.052***	-0.161**
	(0.008)	(0.068)
Exo × Period	0.013	0.122
	(0.015)	(0.086)
Endo × Period	0.006	0.289***
	(0.012)	(0.097)
Observations	5440	4357

Notes: The omitted category "No (constant)" is a binary variable that indicates participation in the "no institution" treatment. The dependent variable in regression (1) is a dummy that equals 1 if contribution is strictly positive and 0 otherwise. The dependent variable in regression (2) is the level of contributions to "PGG Left". (1) Probit regression. (2) Linear regression truncated at 0. Robust standard errors (clustered on part-II groups) in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The effect of the *MCR* on contributions to "PGG Left"

Table A.2: The effect of the *MCR* on contributions to "PGG Left"

	(1) OLS Contribution	(2) OLS Contribution
MCR	0.198***	0.311***
	(0.064)	(0.117)
Period	-0.211***	-0.291***
	(0.039)	(0.048)
No (constant)	11.189***	12.208***
	(0.909)	(1.012)
Exo		-3.687
		(2.736)
Endo		-2.852*
		(1.673)
Exo × Period		0.080
		(0.088)
Endo × Period		0.109
		(0.086)
Observations	5440	5440
Adjusted R^2	0.073	0.078

Notes: The omitted category "No (constant)" is a binary variable that indicates participation in the "no institution" treatment. Robust standard errors (clustered on part-II groups) in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Evidence against imitation

In this part, we test whether our findings are driven by subjects who simply copied their decisions in "PGG Right" to "PGG Left". If that would be the case, then it could explain the treatment effects that we find, because, due to the *MCR*, contributions to "PGG Right" are higher in the "exogenous institution" treatment and the "endogenous institution" treatment than in the "no institution" treatment. As imitation is an intuitive heuristic that doesn't require much thinking, we believe that this would be especially the case for those subjects who submit their contribution decision for "PGG Right" first and let little time pass in between the two decisions.

Averaged over all periods, the proportion of subjects who submitted their contribution decision for "PGG Right" before their decision for "PGG Left" is indeed higher in the "exogenous institution" treatment (34.55 percent) and the "endogenous institution" treatment (37.17 percent) compared to the "no institution" treatment (30.49 percent). Furthermore, in the very first period, the differences are even more pronounced (26.14 percent in the "exogenous institution" treatment and 31.52 percent in the "endogenous institution" treatment vs 15.22 percent in the "no institution" treatment).

However, when we regress the decision to submit first in "PGG Right" on the absolute difference in contributions between "PGG Left" and "PGG Right" (see column (1) of Table A.3), we find that there is no effect in the "no institution" treatment and the "exogenous institution" treatment, but a significantly positive effect in the "endogenous institution" treatment. Thus, submitting the contribution to "PGG Right" first significantly increases the difference in contributions between the "PGG Right" and "PGG Left" in the "endogenous institution" treatment compared to the "no institution" treatment, which is the opposite of what one would expect in case of an imitation effect.

Second, when regressing the absolute difference in time between the submissions of contributions in the two games on the absolute difference in contributions (see column (2) of Table A.3), we don't find any significant effect.

Finally, when regressing all interactions on the absolute difference in contribution between "PGG Left" and "PGG Right" (see column (3) of Table A.3), we find that, in the "no institution" treatment, the difference in contributions significantly decreases the shorter is the difference in time between the two decisions, but only if one decides about "PGG Right" first. Thus, this looks like an imitation effect. However, the effect is completely counteracted in the "exogenous institution" treatment and the "endogenous institution" treatment. Thus, to summarize, we don't find any evidence that imitation effects could explain our results.

Table A.3: Influence of decision sequence and time on the difference in contributions

	(1) OLS Diff. in contr.	(2) OLS Diff. in contr.	(3) OLS Diff. in contr.
Period	0.135***	0.140***	0.138***
	(0.039)	(0.038)	(0.038)
No (constant)	0.563	0.426	0.468
	(0.507)	(0.507)	(0.503)
Exo	5.328***	5.470***	5.306***
	(1.090)	(1.023)	(1.099)
Endo	3.000***	3.447***	2.898***
	(0.919)	(1.070)	(0.993)
Right first	-0.039		-0.781*
	(0.295)		(0.408)
Right first × Exo	0.437		1.279
	(0.830)		(0.896)
Right first × Endo	1.362*		2.182**
	(0.764)		(0.964)
Time diff		0.043	0.034
		(0.043)	(0.036)
Time diff × Exo		-0.009	0.010
		(0.060)	(0.064)
Time diff × Endo		0.008	0.030
		(0.085)	(0.084)
Right first × Time diff			0.641***
			(0.226)
Right first × Time diff × Exo			-0.694***
			(0.244)
Right first × Time diff × Endo			-0.672***
			(0.250)
Observations	5440	5440	5440
Adjusted R^2	0.128	0.125	0.128

Notes: The dependent variable is the absolute difference in contributions between "PGG Left" and "PGG Right". The omitted category "No (constant)" is a binary variable that indicates participation in the "no institution" treatment. "Right first" is a binary variable that is 1 if the subject submitted his decision for "PGG Right" first and zero otherwise. "Time diff" is the time difference, in seconds, between the first and the second submission of contribution decision. Robust standard errors (clustered on part-II groups) in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Voting behavior

Table A.4: Correlations of individual characteristics with voting decisions in period 1

	(1) Vote	(2) Vote	(3) Vote	(4) Vote	(5) Vote	(6) Vote	(7) Vote	(8) Vote	(9) Vote	(10) Vote
Belief about others' cooperativeness (Part I)	0.270***									
	(0.009)									
Unconditional contribution (Part I)		0.237**								
		(0.023)								
Risk			0.047							
			(0.654)							
Patience				0.058						
				(0.581)						
Altruism					0.137					
					(0.190)					
Avg. pos. reciprocity						-0.006				
						(0.949)				
Avg. neg. reciprocity							-0.127			
							(0.226)			
Avg. trust								-0.012		
								(0.913)		
CRT									-0.027	
									(0.832)	
RFT										-0.116
										(0.270)
Observations	92	92	92	92	92	92	92	92	66	92

Notes: Table reports Spearman's rank correlation coefficients. P-values in parentheses. Regression (9) has 66 observations only because we exclude the 26 subjects who indicated that they saw the CRT before. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.5: Effect of control variables on voting decisions in period 1

	(1) OLS Vote	(2) OLS Vote	(3) OLS Vote	(4) OLS Vote	(5) OLS Vote	(6) OLS Vote	(7) OLS Vote	(8) OLS Vote	(9) OLS Vote	(10) OLS Vote
Belief about others' cooperativeness (Part I)	0.347									
	(0.233)									
Unconditional contribution (Part I)		0.257*								
		(0.135)								
Risk			0.409							
			(0.378)							
Patience				0.126						
				(0.236)						
Altruism					1.647*					
					(0.925)					
Avg. pos. reciprocity						0.635				
						(1.022)				
Avg. neg. reciprocity							-0.262			
							(0.400)			
Avg. trust								-0.107		
								(1.895)		

Table A.6: Continuation of Table A.5

	(1) OLS Vote	(2) OLS Vote	(3) OLS Vote	(4) OLS Vote	(5) OLS Vote	(6) OLS Vote	(7) OLS Vote	(8) OLS Vote	(9) OLS Vote	(10) OLS Vote
CRT									-0.261	
									(0.750)	
RFT										-0.047
										(0.052)
Constant	9.863***	10.858***	11.229***	13.162***	8.367**	10.035	14.713***	14.111***	12.976***	14.445***
	(2.869)	(1.788)	(2.575)	(1.429)	(3.319)	(6.286)	(1.520)	(5.126)	(1.090)	(0.972)
Observations	92	92	92	92	92	92	92	92	66	92
R^2	0.041	0.059	0.016	0.003	0.031	0.006	0.005	0.000	0.002	0.009

Notes: Robust standard errors in parentheses. Regression (9) has 66 observations only because we exclude the 26 subjects who indicated that they saw the CRT before. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.7: Probit regression of control variable on voting for a *MCR* of 20 in period 1

	(1) Vote 20	(2) Vote 20	(3) Vote 20	(4) Vote 20	(5) Vote 20	(6) Vote 20	(7) Vote 20	(8) Vote 20	(9) Vote 20	(10) Vote 20
Belief about others' cooperativeness (Part I)	-0.003									
	(0.015)									
Unconditional contribution (Part I)		0.000								
		(0.009)								
Risk			-0.007							
			(0.024)							
Patience				-0.000						
				(0.019)						
Altruism					0.023					
					(0.076)					
Avg. pos. reciprocity						-0.049				
						(0.066)				
Avg. neg. reciprocity							-0.030			
							(0.033)			
Avg. trust								0.114		
								(0.139)		

Table A.8: Continuation of Table A.7

	(1) Vote 20	(2) Vote 20	(3) Vote 20	(4) Vote 20	(5) Vote 20	(6) Vote 20	(7) Vote 20	(8) Vote 20	(9) Vote 20	(10) Vote 20
CRT									0.031	
									(0.052)	
RFT										-0.003
										(0.004)
Observations	92	92	92	92	92	92	92	92	66	92

Notes: Probit regressions. Coefficients report marginal effects calculated at the means. Robust standard errors in parentheses. Regression (9) has 66 observations only because we exclude the 26 subjects who indicated that they saw the CRT before. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Potential channels

Table A.9: Rule-following channel and Social Heuristics Hypothesis

	(1) OLS Contribution	(2) OLS Contribution
No (constant)	11.701***	9.293***
	(1.246)	(1.303)
Exo	2.062	5.321***
	(1.878)	(1.794)
Endo	1.123	2.652
	(1.658)	(1.754)
Period	-0.291***	-0.280***
	(0.048)	(0.061)
Exo × Period	0.080	0.062
	(0.088)	(0.096)
Endo × Period	0.252***	0.210*
	(0.091)	(0.106)
RFT	0.037	
	(0.038)	
Exo × RFT	0.036	
	(0.066)	
Endo × RFT	-0.047	
	(0.059)	
CRT		1.636***
		(0.468)
Exo × CRT		-1.462**
		(0.669)
Endo × CRT		-1.152
		(0.928)
Observations	5440	3900
R^2	0.069	0.080

Notes: The independent variable is contributions to "PGG Left". The omitted category "No (constant)" is a binary variable that indicates participation in the "no institution" treatment. Fewer observation in model (2) as we excluded those subjects that indicated that they had seen the CRT before. Robust standard errors (clustered on part-II groups) in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.10: Votes as a signal of others' cooperativeness (period 1)

	(1) OLS
	Belief
MCR	-0.062
	(0.152)
Vote	0.137
	(0.102)
Avg. vote of others	0.499*
	(0.246)
Constant	3.151
	(3.719)
Observations	92
R^2	0.167

Notes: The independent variable is beliefs about others average contribution to "PGG Left". Robust standard errors (clustered on part-II groups) in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4.A.3 Part I, III, IV - Results

Table A.11: Midpoint of belief interval

	Part I	Part III	Part IV	p-value (I vs. III)	p-value (I vs. IV)	p-value (III vs. IV)
No Inst.	11.88	8.14	9.78	0.000***	0.002***	0.002***
(N=23)	(.38)	(.94)	(.70)			
Exo. Inst.	11.98	11.41	11.74	0.987	0.871	0.291
(N=22)	(.44)	(.78)	(.59)			
Endo. Inst.	11.42	13.02	11.96	0.018**	0.260	0.005***
(N=23)	(.37)	(.66)	(.52)			
p-value (No vs. Exo.)	0.892	0.012**	0.047**			
p-value (No vs. Endo.)	0.422	0.000***	0.013**			
p-value (Exo. vs. Endo.)	0.346	0.196	0.708			

Notes: Numbers in brackets are standard errors. Across treatments: p-values from Wilcoxon rank-sum tests. Within treatments: p-values from Wilcoxon signed-rank tests.

Table A.12: Width of belief interval

	Part I	Part III	Part IV	p-value (I vs. III)	p-value (I vs. IV)	p-value (III vs. IV)
No Inst.	6.37	5.49	6.47	0.055*	0.681	0.004***
(N=23)	(.28)	(.31)	(.24)			
Exo. Inst.	6.35	6.13	6.40	0.464	0.672	0.337
(N=22)	(.23)	(.36)	(.24)			
Endo. Inst.	6.05	5.47	6.74	0.032**	0.011**	0.000***
(N=23)	(.24)	(.35)	(.27)			
p-value (No vs. Exo.)	0.829	0.317	0.900			
p-value (No vs. Endo.)	0.441	0.708	0.574			
p-value (Exo. vs. Endo.)	0.328	0.191	0.374			

Notes: Numbers in brackets are standard errors. Across treatments: p-values from Wilcoxon rank-sum tests. Within treatments: p-values from Wilcoxon signed-rank tests.

Table A.13: Undonditional contributions

	Part I	Part III	Part IV	p-value (I vs. III)	p-value (I vs. IV)	p-value (III vs. IV)
No Inst.	11.37	6.95	8.65	0.000***	0.002***	0.005***
(N=23)	(.60)	(1.01)	(.88)			
Exo. Inst.	11.40	10.40	10.30	0.398	0.168	0.961
(N=22)	(.67)	(.85)	(.75)			
Endo. Inst.	11.53	12.03	11.37	0.553	0.738	0.212
(N=23)	(.61)	(1.05)	(.87)			
p-value (No vs. Exo.)	0.901	0.009***	0.173			
p-value (No vs. Endo.)	0.783	0.002***	0.017**			
p-value (Exo. vs. Endo.)	0.750	0.195	0.159			

Notes: Numbers in brackets are standard errors. Across treatments: p-values from Wilcoxon rank-sum tests. Within treatments: p-values from Wilcoxon signed-rank tests.

Table A.14: Avg. conditional contributions

	Part I	Part III	Part IV	p-value (I vs. III)	p-value (I vs. IV)	p-value (III vs. IV)
No Inst.	7.86	5.63	5.76	0.000***	0.000***	0.867
(N=23)	(.50)	(.54)	(.48)			
Exo. Inst.	8.48	7.60	7.64	0.064*	0.127	0.948
(N=22)	(.45)	(.51)	(.65)			
Endo. Inst.	7.95	7.69	7.20	0.412	0.094*	0.023**
(N=23)	(.35)	(.35)	(.45)			
p-value (No vs. Exo.)	0.414	0.005***	0.011**			
p-value (No vs. Endo.)	0.861	0.008***	0.024**			
p-value (Exo. vs. Endo.)	0.401	0.991	0.829			

Notes: For each subject the average conditional contribution over all possible contributions, from zero to 20, of other group members is calculated and then averaged across part-II groups. Numbers in brackets are standard errors. Across treatments: p-values from Wilcoxon rank-sum tests. Within treatments: p-values from Wilcoxon signed-rank tests.

Table A.15: Treatment effects on preferences for cooperation

	(1) OLS	
	Conditional contribution	
Part I (constant)	2.147***	(0.510)
Part III	-1.078**	(0.415)
Part IV	-1.085**	(0.434)
Exo	-0.696	(0.674)
Endo	-0.613	(0.677)
Part III × Exo	1.327**	(0.635)
Part III × Endo	1.248**	(0.613)
Part IV × Exo	1.779**	(0.714)
Part IV × Endo	1.063*	(0.596)
Contribution of others	0.571***	(0.051)
Part III × Contribution of others	-0.115***	(0.037)
Part IV × Contribution of others	-0.101***	(0.031)
Exo × Contribution of others	0.132*	(0.067)
Endo × Contribution of others	0.070	(0.073)
Exo × Part III × Contribution of others	0.002	(0.068)
Endo × Part III × Contribution of others	0.072	(0.065)
Exo × Part IV × Contribution of others	-0.052	(0.051)
Endo × Part IV × Contribution of others	0.028	(0.069)
Observations	17136	
Adjusted R^2	0.254	

Notes: The omitted category "Part I (constant)" is a binary variable indicating the decision was made in Part I. Robust standard errors (clustered on part-II groups) in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4.A.4 Part V - Results

Table A.16: Individual characteristics

	Risk (0-10)	Patience (0-10)	Altruism (1-4)	Avg. pos. reciprocity (1-7)	Avg. neg. reciprocity (1-7)	Avg. trust (1-4)	CRT (0-3)	RFT (0-30)
No Inst.	5.98	4.82	3.50	5.95	3.17	2.52	1.80	13.72
(N=92)	(.23)	(.28)	(.08)	(.10)	(.14)	(.04)	(.12)	(1.21)
Exo. Inst.	5.70	5.38	3.38	5.97	3.11	2.57	1.76	13.49
(N=88)	(.23)	(.30)	(.08)	(.08)	(.14)	(.04)	(.12)	(1.31)
Endo. Inst.	6.35	5.27	3.32	5.97	3.38	2.66	1.66	13.20
(N=92)	(.19)	(.27)	(.07)	(.08)	(.16)	(.03)	(.12)	(1.23)
p-value (No vs. Exo.)	0.464	0.188	0.140	0.544	0.812	0.490	0.761	0.743
p-value (No vs. Endo.)	0.211	0.258	0.013**	0.455	0.445	0.018**	0.390	0.741
p-value (Exo. vs. Endo.)	0.040**	0.802	0.300	0.938	0.281	0.118	0.570	0.967

Notes: The "Avg. pos. reciprocity", "Avg. neg. reciprocity", and the "Avg. trust" variables are constructed as the average of the answers to three questions. "CRT" is the number of correctly answered questions in the Cognitive Reflection Test. "RFT" is a measure of rule-following propensity. Numbers in brackets are standard errors. p-values are from Wilcoxon ranksum tests.

- Risk question (SOEP, 2004, 2006, 2008, 2009)
  How do you see yourself: Are you generally a person who is fully prepared to take risks or do you try to avoid taking risks?
  Please tick a box on the scale where the value 0 means: "not at all willing to take risks" and the value 10 means: "very willing to take risks".

- Patience question (SOEP, 2008)
  How would you describe yourself: Are you generally an impatient person, or someone who always shows great patience?
  Please tick a box on the scale where the value 0 means: "very impatient" and the value 10 means: "very patient".

- Altruism question (SOEP, 2004, 2008)
  Is it important for you to be there for others?
  Please tick a box on the scale where the value 0 means:"not at all important" and the value 4 means: "very important".

- Reciprocity question (SOEP, 2005)
  For the questions below, please tick a box on the scale, where the value 1 means: "does not apply to me at all" and the value 7 means: "applies to me perfectly".
  (1) If someone does me a favor, I am prepared to return it.
  (2) If I suffer a serious wrong, I will take revenge as soon as possible, no matter what the cost.
  (3) If somebody puts me in a difficult position, I will do the same to him/her.
  (4) I go out of the way to help somebody who has been kind to me before.
  (5) If somebody offends me, I will offend him/her back.
  (6) I am ready to undergo personal costs to help somebody who helped me before.

- Trust question (SOEP, 2003, 2008)
  For the questions below, please tick a box on the scale, where the value 1 means: "totally disagree" and the value 4 means: "totally agree".
  On the whole one can trust people.
  Nowadays one can't rely on anyone.
  If one is dealing with strangers, it is better to be careful before one can trust them.

B.1.1 Instructions for Part I

General instructions

Welcome to this experiment.

Please read this first page of the instructions carefully on your own. We will then read the rest of the instructions aloud in front of all participants.

In the experiment you can earn a considerable amount of money in addition to the **10 Swiss francs** that you receive for showing up on time. How much you earn will depend on your own decisions and those of the other participants. It is thus very important that you read the instructions carefully. If you have any questions please do not ask aloud but raise your hand.

During the experiment, speaking with the other participants and the use of mobile phones are not allowed. Violation of these rules can lead to exclusion from the experiment and loss of all associated earnings.

During the experiment, we will refer to earnings in Experimental Currency Units, or ECU. Your entire income will first be calculated in ECU. The ECU you earn during the experiment will be converted to Swiss francs at the end of the experiment, according to the following conversion rate:

100 ECU = 3 CHF

At the end of today's experiment, you will receive these earnings plus the show-up payment of 10 Swiss francs **in cash**.

At no point, during or after the experiment, will you learn the identities of the people with whom you interact during the experiment, nor will these people learn your identity.

The Experiment

The experiment consists of several parts. At the beginning of each part you will receive instructions that explain that part of the experiment. The earnings that you will receive for the experiment consist of the sum of your earnings in the individual parts in addition to the fee for showing up.

Part I

Group Membership

At the beginning of Part I, the computer will assign you at random to a group consisting of **four participants**. All interactions during Part I take place within the group to which you are assigned.

Part I consists of two phases. In both phases you will make decisions related to a basic task. Before explaining the two phases in more detail, we first explain the basic task to you.

The Basic Task

Each of the four members of your group is endowed with 20 tokens. Each member of the group decides how many of the 20 tokens to put in a private account and how many to contribute to a group account. Any tokens you put in the private account cannot be contributed to the group account and vice versa. You can earn income from the private account as well as from the group account.

Your income from the private account

For each token you put in your private account you earn an income of one ECU. Nobody except you earns anything from tokens you put in your private account.

EXAMPLE: If you put 6 tokens in your private account, you earn 6 ECU from the private account.

Your income from the group account

For each token you contribute to the group account you and the other three group members each receive 0.5 ECU. Note that you will also earn income from the tokens that other group members contribute to the group account. For each group member the income from the group account will be determined as follows:

Each group member's income from the group account

= 0.5 * sum of all tokens contributed to the group account

Put differently, the total number of tokens in the group account will be doubled and then equally distributed among all four group members. This yields, for each group member, 0.5 times the total number of tokens contributed. Suppose you contribute one token to the group account. The sum of tokens in the group account would then rise by one token. Your income from the group account would, thus, rise by 0.5 * 1 = 0.5 ECU. The income of each other group member would also rise by 0.5 ECU. So, contributing one token to the group account generates total income for the group of 4 * 0.5 ECU = 2 ECU.

*EXAMPLE: If the sum of tokens in the group account is 60 tokens, then you and all other group members each earn an income of 0.5 * 60 = 30 ECU from the group account. The total income for the group from the group account is 4 * 30 ECU = 120 ECU.*

Your total income

Your total income equals the sum of your income from the private account and your income from the group account.

Total income =

Income from the private account + Income from the group account =

(20 – tokens you contribute to the group account) + (0.5 * sum of tokens in the group account)

Part I of the experiment consists of two phases. In the first phase you are asked to indicate your belief about how much **the other three members** of your group will, on average, contribute to the group account in a task identical to the one just described. In the second phase you are asked to decide how much **you** contribute to the group account in a task identical to the one just described.

a) Phase 1: Estimates of other group members' average contributions

In Phase 1, we ask you to estimate how many tokens the other three members in your group will, on average, contribute to the group account in Phase 2. Remember that each member has an endowment of 20 tokens and can contribute any amount from 0 to 20 tokens to the group account. Specifically, we ask you to provide **a range of values** that you believe will contain the average number of tokens that the other group members contribute to the group account. You will enter your estimate as two integers: one number for the lower end of the range and another for the higher end of the range.

In Phase 2 of Part I, all group members will decide how much to actually contribute to the group account. We will round the actual average contribution of the other group members to the nearest integer, and compare it to the range you specified. You will earn ECU only if the actual (rounded) average contribution of others to the group account lies inside the range you specify. Furthermore, the wider the range you indicate the smaller are your potential earnings. More precisely, the exact amount you earn is calculated according to the following rules:

- If the actual (rounded) average lies outside of the range you specify you earn 0 ECU.
- The maximum you can earn is 20 ECU. You earn 20 ECU if you (a) specify only a single value – that is, if the lower number and the higher number you specify are the same – and (b) this value is equal to the actual (rounded) average contribution of others to the group account. So, for example, if you are certain that the average contribution of others will be 15 then you should enter 15 for both the lower number and the higher number. If the (rounded) average of others is actually 15 you will earn 20 ECU.
- As the range you specify becomes wider, you earn **less** money for a correct estimate. Specifically, for every unit that your range increases in width, your potential income decreases by 1 ECU. So, for example, if you enter 8 for the lower and 20 for the higher end of the range (i.e. your range has a width of 12) and the actual (rounded) average contribution of others is 14 tokens, then you will earn 20 – 12 = 8 ECU. You earn more than you would earn if you had entered a wider range, say from 5 to 20 (income 20 – 15 = 5 ECU), but you earn less than you would earn if you had entered a narrower range, say from 10 to 15 (income 20 – 5 = 15 ECU) or if you had entered a range consisting only of 14 (income = 20 – 0 = 20 ECU).

If you enter 0 for the lower end and 20 for the higher end, your range covers all possible average token amounts and the actual (rounded) average of others' contributions is thus guaranteed to lie in your range. In this case, you earn nothing (income = 20 – 20 = 0 ECU).

To summarize, the rule is that you earn money for specifying a range that contains the actual average of others' contributions,, but the amount you earn for such a correct estimate is smaller the wider the range you indicate.

b) Phase 2: "Unconditional contribution" and "contribution table"

In the second phase of Part I, you will decide about your actual contribution to the group account. You have 20 tokens and you can choose to contribute any of these tokens to a group account. The tokens you do not contribute are put into your private account.

In this phase, you will make two types of contribution decisions: an **unconditional contribution** decision and a decision through a **contribution table**. Only one of these decisions will count, but you will not know which one until the end of the experiment. This means that you should treat each one as if it is the one that determines your earnings from Phase 2.

- In the **unconditional contribution** decision, you decide how many of the 20 tokens you contribute to the group account. You will enter your contribution decision as a single number between 0 and 20.

- In the decision through a **contribution table** you may contribute different amounts for each possible average unconditional contribution of the other group members (rounded to the nearest integer). That is, you have to specify how much you want to contribute if the other three group members contribute, on average, 0 tokens, 1 token, 2 tokens, etc., up to 20 tokens, to the group account. You will see a table, with all 21 possible integer values from 0 to 20, corresponding to the possible average unconditional contributions made by the other three group members.

Earnings from Part I

After all four participants in a group have made both types of decisions in Phase 2, your earnings from Part I will be determined as follows.

- First, the computer will compare the range you provided as an estimate for the other group members' average contributions to their actual average **unconditional contributions**. This will determine your earnings from Phase 1.
- Second, the computer will randomly select three group members to have their unconditional contributions count as their contribution decision. The computer will then calculate the average unconditional contribution of the three selected group members. This average determines how much the remaining group member will contribute, based on how that group member completed the contribution table. Together this determines the actual contributions of all four group members and, thus, each member's earnings from Phase 2.

*EXAMPLE: Assume that the three group members that were randomly selected to have their unconditional contributions count decided to contribute 0, 3, and 15 tokens. The average contribution of these three group members, therefore, is 18/3 = 6 tokens. The computer will then check the contribution table of the remaining group member, for the entry in the row corresponding to an average contribution of 6, and will use this entry to determine the contribution decision of this fourth group member. Suppose that this group member decided to contribute 10 when the average contribution by other group members is 6. Then, the computer will make this group member contribute 10. The total sum of contributions to the group account is thus 0 + 3 + 15 + 10 = 28 tokens. All group members, therefore, earn 0.5 * 28 = 14 ECU from the group account* ***plus*** *their respective incomes from the private accounts.*

You will make these decisions only once in Part I. You will be informed about the contribution decisions of the other group members and your payoff from Part I at the end of the experiment, after everyone has made all decisions in the experiment.

Do you have any questions? If yes, please raise your hand. We will then come to you at your workplace. If not, please click "Continue" on your computer screen.

Once we have answered all questions, we will ask you to answer some comprehension questions on your computer screen. These questions will ensure that everyone understands the instructions for Part I.

B.1.2 Decision screens for Part I

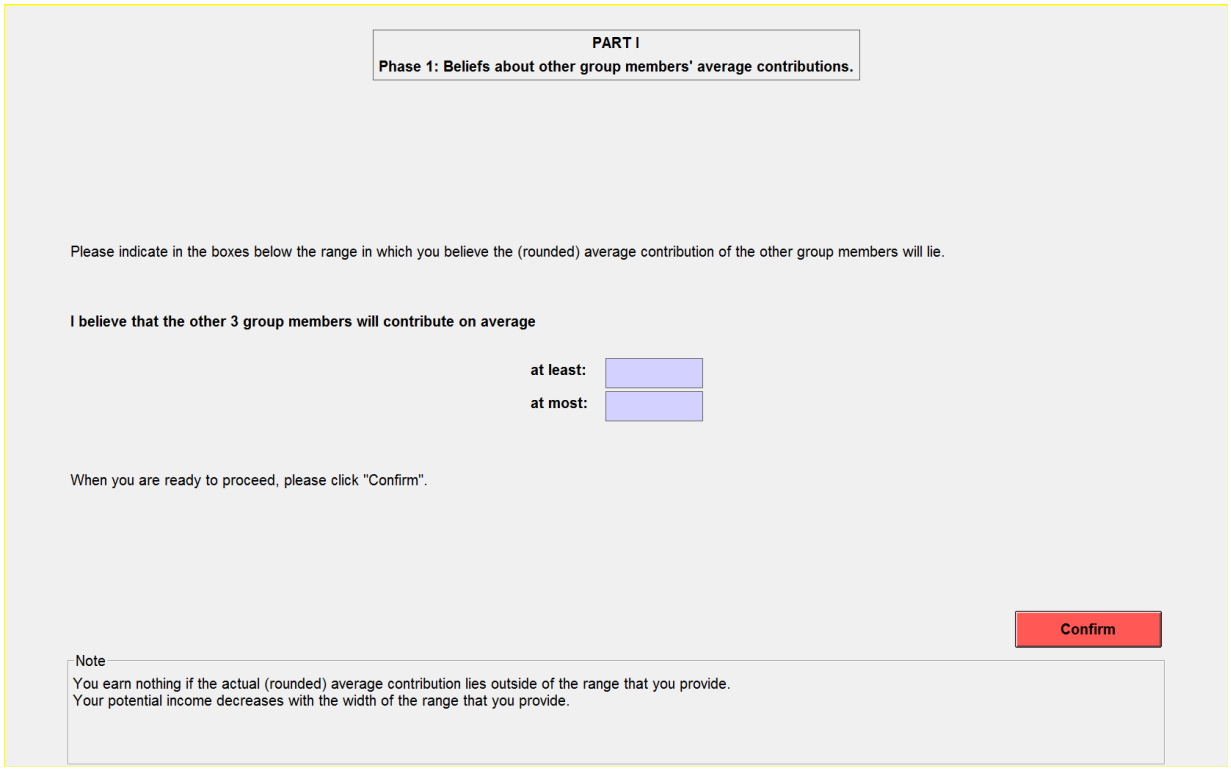


Figure B.1: Beliefs about others' contributions.

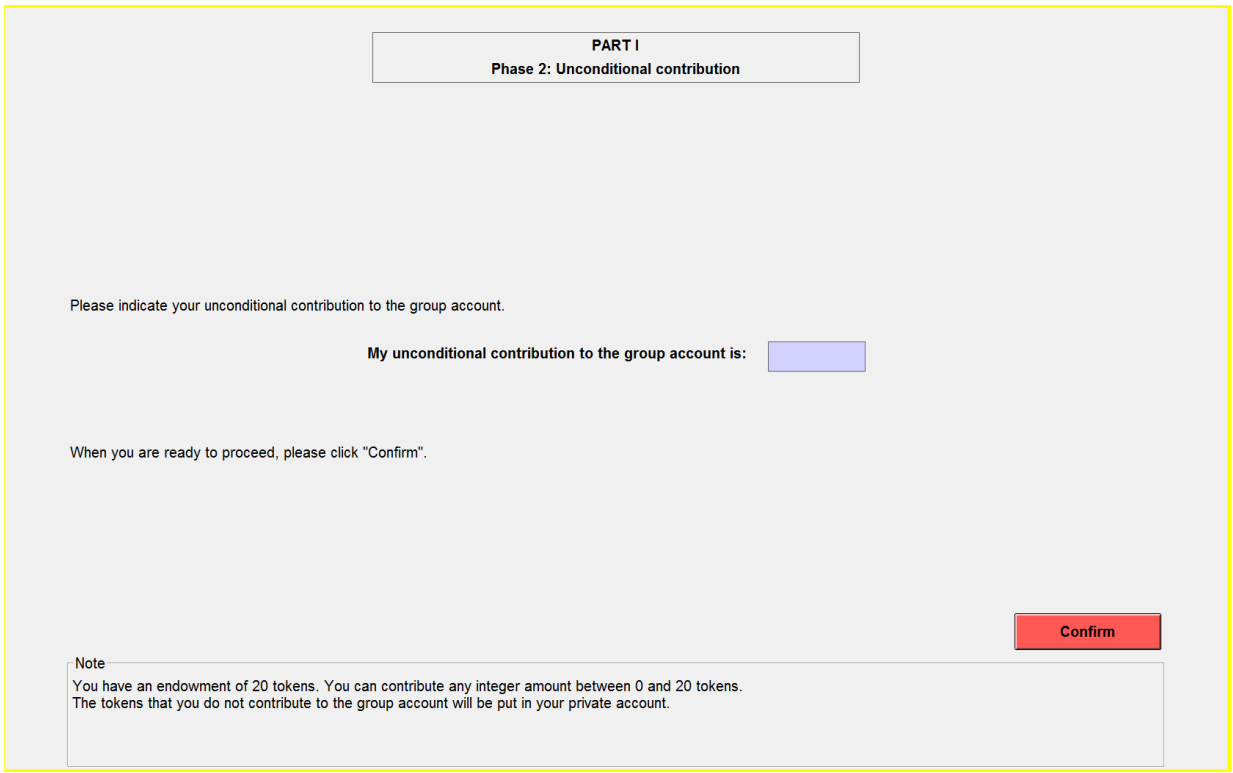


Figure B.2: Unconditional contribution decision.

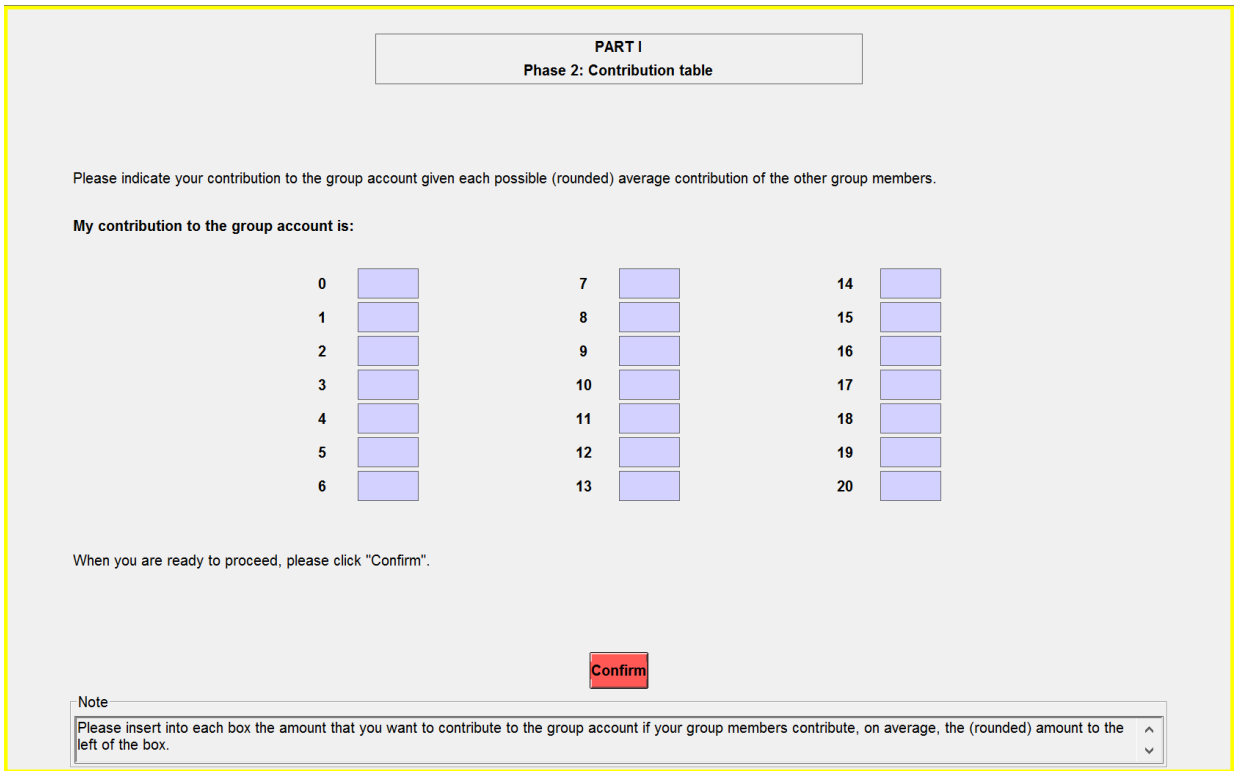


Figure B.3: Conditional contribution decision.

B.2.1 Instructions for Part II - "Endogenous institution" treatment

Part II

Group membership

At the beginning of Part II, the computer will assign you at random to a group consisting of four participants **that you have not interacted with before**. This part of the experiment consists of 20 periods and all interactions during Part II take place **with the same group members**. In each period, you will simultaneously participate in two tasks. They will be displayed next to each other on the same computer screen and we will, thus, refer to these as **Task Left** and **Task Right**.

For each task, you have a separate endowment of 20 tokens that you can contribute to a group account or put in your private account, similar to the basic task in Part I. In Part II, everyone will make unconditional contributions. You will enter, separately, the number of tokens you decide to contribute to the group account in Task Left and in Task Right.

Task Left

On the left side of the computer screen, you will decide how many of your endowment of 20 tokens to contribute to the group account and how many to put in your private account. You can enter any integer from 0 to 20. Your income from Task Left is calculated in the same way as described for the basic task and, thus, depends on your contribution and the contributions of the other three members of your group.

Task Right

On the right side of the computer screen, your group will, at the beginning of each period (and thus before a decision in Task Left can be made), first vote on a "**contribution threshold.**" The contribution threshold specifies a minimum level of contribution to the group account in Task Right for each group member. The contribution threshold can be any value between 0 and 20.

The contribution threshold affects the income of group members from Task Right, depending on whether they contribute at least as many or fewer tokens to the group account than specified by the contribution threshold. Specifically:

- The income from Task Right of any group member who contributes **at least as many** tokens to the group account as specified by the contribution threshold is **not affected** by the contribution threshold. The income from Task Right is then determined as described for the basic task.

- Any group member who contributes **fewer** tokens to the group account than the minimum level specified by the contribution threshold **loses any income** from Task Right. That is, a group member that contributes less than the contribution threshold receives an income of 0 for Task Right, regardless of how much this group member or other group members contributed. Thus, there is a penalty for contributing fewer tokens to the group account than the contribution threshold, and the penalty is the loss of all income for that period in Task Right. **A participant's income in Task Left is not affected by anything that happens in Task Right and vice versa.** Similarly, if one participant is penalized for contributing less than the contribution threshold in Task Right, the incomes of other participants are not affected. Thus, the other group members still benefit from any contributions made by any group member in Task Right.

*EXAMPLE: The contribution threshold is set to 15 in Task Right. Group member A contributes 5, member B 15, member C 20, and member D 20 tokens to the group account in Task Right. The total contributions are thus 60 tokens. Member A earns 0 ECU from Task Right, because he contributed less than the "contribution threshold" of 15 tokens. Member B earns 5 ECU from the private account plus an income of 0.5 * 60 = 30 ECU from the group account from Task Right. Member C and Member D earn 0 ECU from the private account plus 30* ECU *from the group account in Task Right. Note that all group members also earn money based on what happens in Task Left, which is independent of Task Right.*

How the contribution threshold for Task Right is determined:

At the beginning of every period, before any contribution decisions are made, all four members of a group vote on the contribution threshold for Task Right for that period. Every member votes for a desired contribution threshold, by specifying an integer value between 0 and 20.

The implemented contribution threshold for Task Right for that period is the <u>lowest value</u> voted for by any group member.

EXAMPLE: Assume that group member A votes for 7, group member B for 12, group member C for 18, and group member D for 10. The implemented contribution threshold for Task Right in that period is 7, the lowest vote in the group. Any group member who contributes less than 7 tokens in that period in Task Right then earns 0 ECU from Task Right.

After the voting takes place, all group members are informed about the implemented contribution threshold and about all of the separate votes cast by members of the group. The votes will be presented in descending order and it is not possible to identify which member of the group voted for which value of the contribution threshold.

Before you make your contribution decisions in Task Left and Task Right, we will ask you about your belief about the other group members' average contribution in Task Left and Task Right. Contrary to Part I, you will enter your (rounded) belief as a single number. So, for example, if you believe that the (rounded) average contribution is 12 in Task Left and 8 in Task Right, you should enter the numbers 12 and 8 in the respective input boxes on the screen. Whether your beliefs are correct or not does not impact your payoff. Please enter your best estimates.

After that you will make your contribution decisions in Task Left and Task Right.

Summary

You will make the following decisions in Part II:

- You will vote on a contribution threshold for Task Right. The contribution threshold changes the potential payoffs only in Task Right. It has no effect on the payoffs from Task Left.
- You will enter your beliefs about the average contribution of the other group members in Task Left and Task Right.
- You will then make two contribution decisions, one in Task Left and one in Task Right.

Total income in each period

Your total income in each period is equal to the sum of your incomes in the two tasks. So, for example, if you earn 30 ECU from Task Left and 10 ECU from Task Right, your total income in that period will be 40 ECU. At the end of each period all group members will be informed about their incomes in Task Left and Task Right and the respective contributions of all group members. The contributions will be presented in descending order and it is not possible to identify which member of the group contributed which number of tokens to the group accounts in Task Left and Task Right.

Earnings from Part II

At the end of the experiment, one out of the twenty periods from Part II will be randomly selected to count for payment. Your decisions and those of your group members in that period will then be implemented and will determine your earnings from Part II. Specifically, **your payoff for the randomly selected period will be multiplied by 20, so that it counts for all 20 periods.** Note that every decision in each of the twenty periods can be relevant for your payoff. It is therefore important that you make your decisions in every period as if it would be the period that determines your actual payoff.

Do you have any questions? If yes, please raise your hand. We will then come to you at your workplace. Once we have answered all questions, we will ask you to answer some comprehension questions on your computer screen. These questions will ensure that everyone understands the instructions for Part II.

B.2.2 Decision screens for Part II - "Endogenous institution" treatment

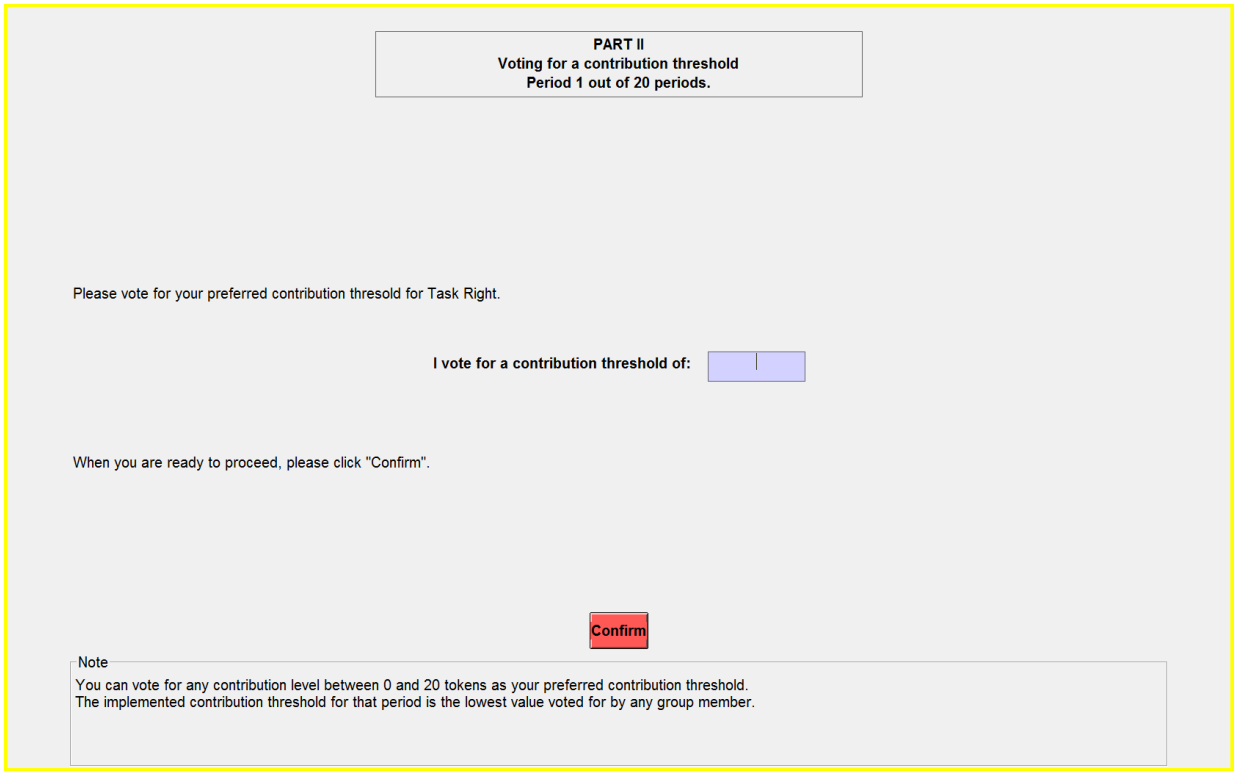


Figure B.4: Voting decision.

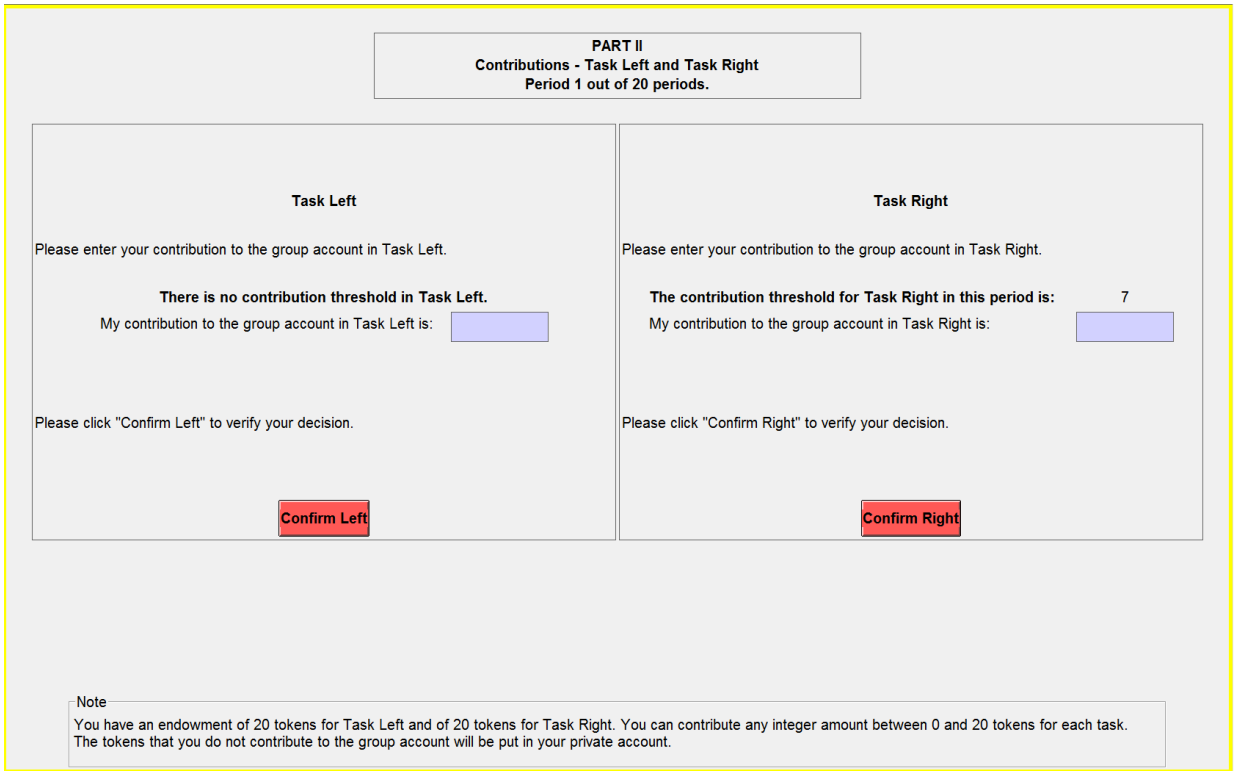


Figure B.5: Contribution decision for "PGG Left" and "PGG Right".